2024 Neuro Inspired Computational Elements Conference (NICE) | 979-8-3503-9058-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/NICE61972.2024.10548306

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 Electrical and Computer Engineering
University of California, Santa Barbara
California, USA
†saisukruthbezugam@ieee.org, ‡strukov@ece.ucsb.edu, §bongjin@ucsb.edu

***Abstract*—In this study, we propose the first hardware implementation of a context-based recurrent spiking neural network (RSNN) emphasizing on integrating dual information streams within the neocortical pyramidal neurons specifically Context-Dependent Leaky Integrate and Fire (CLIF) neuron models, essential element in RSNN. We present a quantized version of the CLIF neuron (qCLIF), developed through a hardware-software codesign approach utilizing the sparse activity of RSNN. Implemented in a 45nm technology node, the qCLIF is compact (900um²) and achieves a high accuracy of 90% despite 8 bit quantization on DVS gesture classification dataset. Our analysis spans a network configuration from 10 to 200 qCLIF neurons, supporting up to 82k synapses within a 1.86 mm² footprint, demonstrating scalability and efficiency.**

***Index Terms*—Spiking neural network accelerator, hardware software codesign, neocortical neurons, CLIF neurons**

I. Introduction

As the demand for more efficient and capable computing systems grows, neuromorphic computing has emerged as a promising avenue for emulating brain-like processing capabilities. This field, bridging artificial intelligence and neuroscience, not only aims to replicate human brain functions but also seeks to drastically reduce the power consumption of computational systems [1], [2]. In this paper, we explore how the integration of advanced neuron models can potentially address these challenges. Spiking Neural Networks (SNNs), termed the third generation of neural networks, offer a pathway to this goal through spike-based computations. However, many SNNs have not yet achieved the accuracy levels of ANNs. Efforts to bridge the performance gap have explored various approaches, including exploiting the multi-timescale dynamics of neurons [10], [11], [23], local learning algorithms [12], [13]. A notable advancement in this domain is the integration of context-dependent leaky integrate and fire (CLIF) neurons into recurrent spiking neural networks (RSNNs) [3], [14].

The fundamental concept posits that understanding or locating items becomes more manageable with appropriate context. Consider the scenario of entering a disorganized room and being asked to find an object; the task proves challenging, overwhelmed by numerous options, making decision-making difficult. However, if the request specifies context, such as "find something to play music with," the search simplifies due to the targeted nature of the inquiry. This process involves two distinct streams of information: 1) stimuli information and 2) context. Though independent, correlating these streams enhances object identification efficiency. Drawing parallels with neocortical pyramidal neurons, this method employs dual information pathways: the bottom-up stimuli received by basal dendrites and the top-down context provided to apical tufts. Each pathway processes separate information—basal dendrites handle stimuli while apical tufts manage context. A correlation

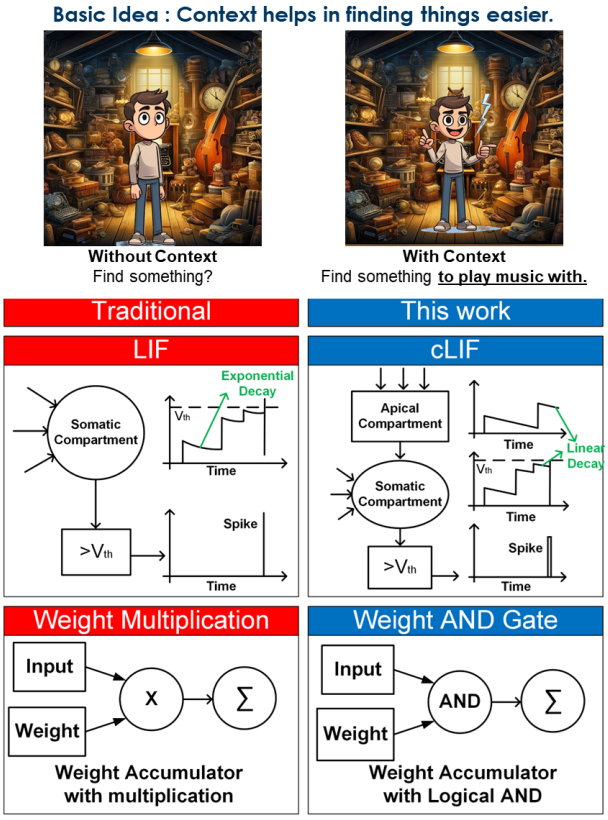

Fig. 1. Top: Basic idea of using context. Bottom : Comparison of Neuronal Models and Synaptic Processing: Traditional LIF versus proposed qCLIF. The Traditional LIF model exhibits a single-compartment with exponential decay and spike generation upon reaching threshold voltage, with an analog weight multiplication approach. The qCLIF model introduces an additional apical compartment with linear decay dynamics and utilizes a digital weight AND gate mechanism for synaptic processing, offering advantages in speed, reconfigurability, robustness, and scalability of digital hardware methodology.

*Equal Contribution

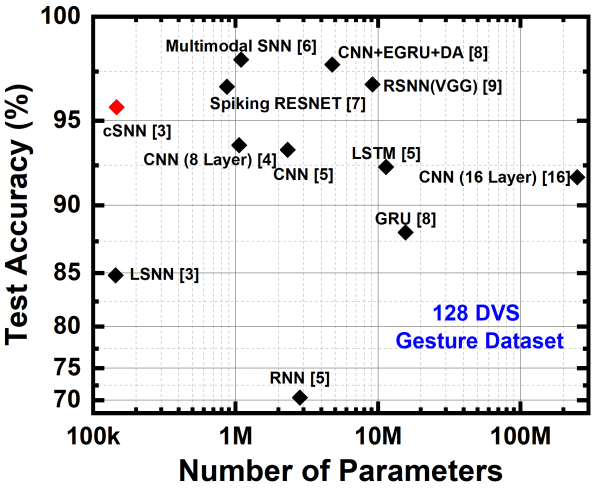

Fig. 2. A comparison of test accuracy versus parameter count for diverse network architectures addressing the DVS Gesture dataset. Notably, context based Recurrent Spiking Neural Networks (cSNNs) achieve high accuracy with significantly fewer parameters compared to other models referenced in the literature.

between these streams may trigger an higher output frequency. This integration has improved accuracy in gesture classification and speech recognition tasks, outperforming several existing models [4]–[9].

The unique feature of CLIF neurons in RSNNs is their use of contextual input to enhance the somatic compartment's computational capacity (see Fig. 1). Using this additional stream of information as context these networks give accuracy on par with most of the network which are much larger in size Fig. 2. There have been no hardware implementations of the CLIF neuron model, although analog and digital versions of other neuron models exist. This work introduces a hardware-friendly variant of the CLIF neuron: the Quantized CLIF neuron. This model retains the accuracy of the original while being more amenable to digital implementation. The study systematically analyzes the model's performance with quantized neuron parameters and weights and examines network activity patterns.

The proposed neuron model, including its synaptic compartment, was implemented using open source 45 nm technology [15]. Layout and post-synthesis evaluations assessed area, power, and timing characteristics. These assessments provide critical insights into the feasibility and efficiency of implementing the qCLIF neuron model in neuromorphic hardware. The following sections will elaborate on the detailed methodology, results, comprehensive qCLIF neuron model analysis, hardware synthesis, and performance evaluations.

II. Proposed Quantized Context-Based Leaky Integrate and Fire Neuron Model

A. Mathematical Model

Neuronal computation is segregated into two primary domains: the apical compartment, which assimilates contextual information, and the somatic compartment, which processes the primary stimulus inputs like spikes from various sensory modalities. The classical CLIF neuron model [3] captures this bifurcation with equations that reflect the dynamic interplay between these compartments, as shown in eq. (1) (2).

$$V_j^a(t+\Delta t) = \alpha V_j^a(t) + (1-\alpha) R_m I_j^a(t+\Delta t) \tag{1}$$

$$\begin{aligned} V_j(t+\Delta t) =& \beta V_j(t) + (1-\beta)[R_m I_j(t+\Delta t)\cdot \\ & \text{ReLU}(V_j^a(t+\Delta t))] - V_{th} \end{aligned} \tag{2}$$

where $\alpha = \exp\left(-\frac{\Delta t}{\tau_a}\right)$ and $\beta = \exp\left(-\frac{\Delta t}{\tau_m}\right)$ are the exponential decay constants for the apical and somatic potentials, respectively. Δt is typically set at 1 ms, akin to biological neurons. I_j^a and I_j represent apical and somatic (stimuli) input currents. R_m denotes the membrane resistance, V_{th} is the spiking threshold, and $s_j(t)$, which can be either 1 (indicating a spike) or 0 (no spike) V_j greater than or less than V_{th} respectively. To tailor these dynamics for digital systems, we adjust Δt to equate to a single simulation timestep or clock cycle, which aligns the model with the discrete nature of digital computation. Further, we optimize by approximating many computational steps. The proposed qCLIF neuron model is expressed in eq. (3) (4).

$$V^{ap}(t+dt) = V^{ap}(t) - \alpha_{\text{leak}} + V_{\text{input}}^{ap}(t+dt) \tag{3}$$

$$\begin{aligned} V^{\text{som}}(t+dt) =& V^{\text{som}}(t) - \beta_{\text{leak}} \\ & + \left(\text{RELU}\left(V^{ap}(t+dt)\right) \cdot V_{\text{input}}^{\text{som}}(t+dt)\right) \end{aligned} \tag{4}$$

α_{leak} and β_{leak} are the linear decay constants. $V_{\text{input}}^{\text{ap}}(t+dt)$ and $V_{\text{input}}^{\text{som}}(t+dt)$ are the contextual and stimulus inputs chosen to be of 'N' bit-width fixed-point numbers, respectively. A new issue arises with constant linear leakage: it can cause compartment voltages to fall below zero uncontrollably. To address this, we set the lower limit of the voltages to zero. Upon neuron spiking, a 'Reset to Zero' mechanism is applied to the somatic compartment. These inputs are defined by eq. (5) (6).

$$V_{\text{input}}^{\text{con}}(t+dt) = \sum(\text{AND}(C_{\text{spike}}, W_{\text{context}})) \tag{5}$$

$$\begin{aligned} V_{\text{input}}^{\text{som}}(t+dt) =& \sum(\text{AND}(S_{\text{spike}}, W_{\text{soma}})) \\ & + \sum(\text{AND}(P_{\text{spike}}, W_{\text{recurrent}})) \end{aligned} \tag{6}$$

C_{spike} and W_{spike} represent the contextual and somatic spike inputs, W_{context}, W_{soma}, and $W_{\text{recurrent}}$ are the corresponding synaptic weights, and P_{spike} denotes the previous spike information for recurrent connections. Despite these modifications, the fundamental characteristics of neuronal activity are retained. The model leverages a piecewise-linear approach to mimic the neuron's response, especially in operational ranges where linear and exponential decay patterns are virtually

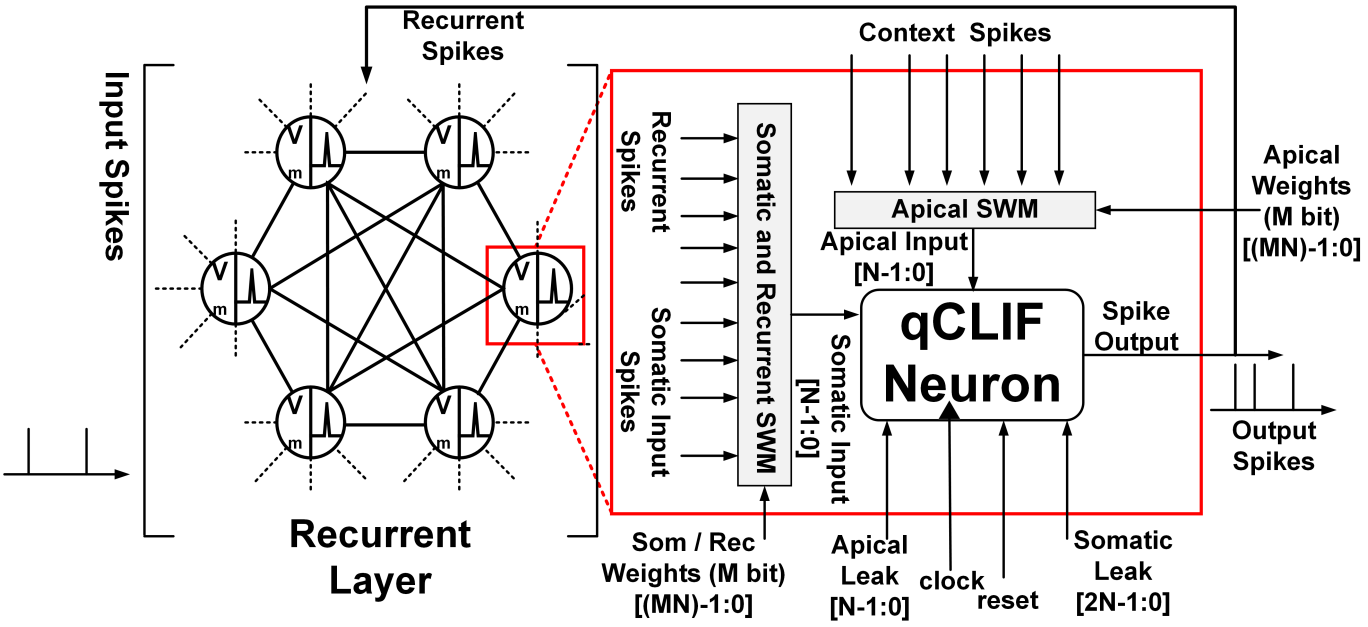

Fig. 3. Architecture of a Recurrent Spiking Neural Network layer made of qCLIF Neurons. The inset view highlights a single qCLIF neuron, which processes both input and recurrent spikes using a combination of somatic and apical inputs along with somatic and recurrent, apical weights. It operates with a set of parameters including somatic leak and apical leak, governed by a clock and reset mechanism generating a train of output spikes.

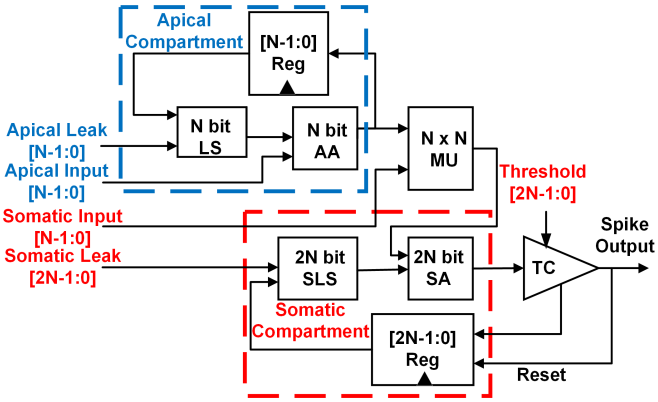

Fig. 4. Digital design of N bit qCLIF neuron. LS : Leakage Subtractor, AA: Apical Accumulator, MU: Multiplication Unit, SLS: Somatic Leakage Subtractor, SA: Somatic Accumulator, TC: Threshold Comparator.

indistinguishable. Moreover, the model effectively captures the core processes of neuronal dynamics: it integrates and decays inputs within the apical and somatic compartments and enhances the somatic potential through interaction with the modulated apical input. The next section will discuss the digital implementation of the proposed qCLIF model.

B. Architecture

For building a recurrent layer of qCLIF neurons, a modular approach is employed to process and integrate both external inputs and internally generated feedback, which is a key characteristic of the dynamics in recurrent neural network systems. This architecture ensures efficient spike processing and temporal data integration. The proposed architecture is outlined in Fig. 3 and digital design of qCLIF is shown in Fig. 4. The key modules in this architecture are detailed as follows: **Spike Weighting Module (SWM):** The SWM processes incoming spikes from external somatic and apical stimuli. Additionally, it handles spikes generated internally by the network, which are used as inputs in the next step. Incorporating recurrent feedback is crucial for the temporal dynamics inherent in the network's processing. Each external or recurrent spike is combined with its corresponding M-bit weight value using bitwise AND operations. The module employs Carry Save-Ahead (CSA) adders optimized for speed with a pipeline structured in $log_3(N)$ stages. **Apical Compartment (AC):** This component consists of a Leakage Subtractor (LS) and an Apical Accumulator (AA). The AC processes the outputs from the SWM, with the LS managing linear leakage from the accumulated data and the AA adding contextual inputs along with the recurrent feedback. Output registers connected to the AA store the cumulative sums, allowing for their reuse in subsequent accumulation cycles. The signed bit is checked in each clock cycle. Whenever the output is negative, the register is reset to 0. **Multiplication Unit (MU):** The MU receives combined outputs from the AC, including external and internal data. It multiplies this data with additional stimuli inputs using an array multiplier, producing a $2N$-bit binary product that is then sent to the Somatic Compartment for further processing. **Somatic Compartment (SC):** Structurally similar to the AC, the SC comprises the Somatic Leakage Subtractor (SLS) and the Somatic Accumulator (SA). It accumulates the $2N$-bit data from the MU, with the SLS adjusting for leakage and the SA summing the inputs for threshold comparison. **Threshold Comparator (TC):** In the final stage, the TC compares

the aggregate output from the SC against a predetermined threshold. If the output surpasses this threshold, a spike is generated. This output spike plays a dual role as both the neuron's output and an input for the SWM in the subsequent computational cycle, perpetuating the recurrent feedback loop within the network. Further, the spike is connected as *RESET* to somatic compartment register.

III. RESULTS

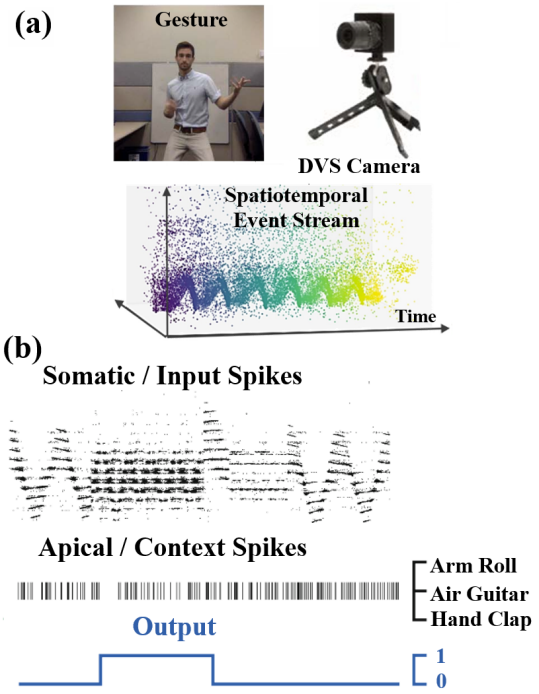

Fig. 5. (a) Setup for Gesture Recognition Using DVS: This setup is utilized for recording gestures, which generate spatio-temporal event streams. The images are adapted from source [16]. (b) Simulation Setup: This involves the use of a spatio-temporal stream as somatic input spikes. Context spikes are directed to the apical compartment, prompting the network to recognize a specific class. The output indicates whether the input stream corresponds with the context spikes.

The efficacy of the proposed qCLIF neuron model was evaluated within a context-dependent RSNN using the Dynamic Vision Sensor (DVS) Gesture dataset [16]. This dataset comprises ten distinct categories of hand gestures, each recorded with a spiking vision sensor. Inputs to the model are structured as 512-dimensional spike trains, with durations ranging from 196 to 1476 milliseconds. The network architecture consists of 200 cLIF neurons, connected recurrently through their somatic compartments. A context input mechanism incorporating ten neurons, each aligned with a specific class, was integrated. The target class is indicated by Poisson spikes at 200 Hz from the respective neuron, and the network's output is binary, indicating the correspondence of a gesture to the target class. The simulation setup The training was conducted over ten epochs using Backpropagation Through Time (BPTT) with an Adam optimizer. In Fig. 5, we present both the practical setup for gesture recognition using a DVS and the corresponding

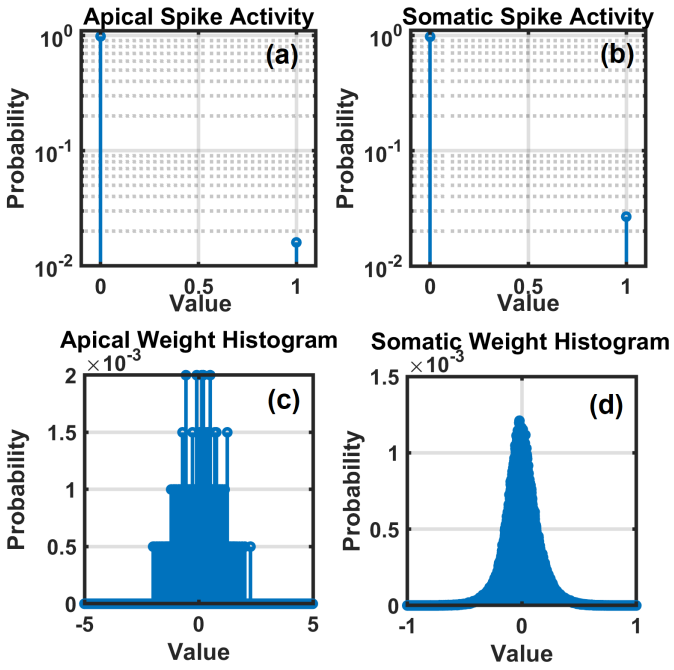

Fig. 6. Insights from the network simulation (a) Apical input spike activity histogram, (b) Somatic input spike activity histogram, (c) Apical weight Histogram, (d) Somatic weight histogram.

simulation framework, illustrating how spatio-temporal event streams are recorded and processed for gesture classification. For this task, we have determined that both somatic leak and apical leak are uniformly distributed across neurons, with values of (200, 7) and 7 respectively. This distribution is aimed at aligning their exponential decay parameters at 200 δt and 20 δt. Additionally, we have adopted other parameters from [3].

Initial tests focused on the impact of linear leakage of qCLIF on network performance. With optimized leakage parameters (choosing a linear regime of exponential decay), the qCLIF model's performance was found to be slightly below the state-of-the-art accuracy of 95% by 0.5%, proving its efficacy. Furthermore, removing leakage from the neuron model led to a notable 4% decrease in accuracy at full precision.

The potential of quantization was also examined through fixed point numbers for constants and variables in eq. (5) (6). Considering over 500 inputs per neuron (context spikes, stimuli spikes, and recurrent spikes) processed through associated weights (assumed quantized to 8 bits), an adder precision of at least 16 bits was deemed necessary (to handle a maximum total of 127,500). This would necessitate a 16x16 multiplier and a final 32-bit adder. However, an in-depth analysis of network activity showed that spiking occurred in only about 2% of the neuron population, as depicted in Fig. 6 (a-b). This finding justified the use of smaller bit widths.

The effects of quantizing neurons and weights on network performance were also examined. The weight distribution within the network was observed to follow a normal distribution, with soma weights predominantly in the range of -0.5

TABLE I
EFFECT OF QUANTIZATION ON NETWORK PERFORMANCE

Precision Level	Neuron Quantization Accuracy (%)	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

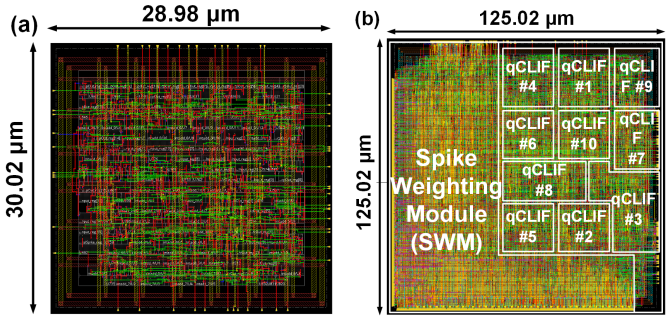

Fig. 7. Layout of the proposed design: (a) single cLIF neuron, (b) a complete 10 neuron qCLIF model with weight accumulator.

to 0.5 and apical weights between -2 and 2. This informed the decision to quantize weights within these specific ranges, deviating from the common -1 to 1 range, as shown in Fig. 6 (c-d). This tailored approach to quantization aimed to fully utilize the range, thus enhancing network efficiency and effectiveness. The consequent impact of quantization on model performance is elaborated in Table I.

The proposed qCLIF design, synthesized on a 45nm CMOS process [15], was realized through Synopsys Design Vision and Innovus automation tools. Simulation results for a single neuron with 8-bit precision indicated an area footprint of 0.029 × 0.030 mm² (layout in Fig. 7 (a), a slack time of 5.62 ns, and power consumption metrics as follows: switching power at 0.020 mW, internal power at 0.041 mW, and leakage power at 0.016 mW. This yielded a total power consumption of 0.077 mW and an energy efficiency of 0.773 pJ per spike, as detailed in Table II. The performance metrics provided in this study primarily reflect worst-case scenarios and may not fully capture variations specific to different activities. It's plausible the networks exhibit sparse activity, hence could potentially demonstrate lower energy consumption. Fig. 7 (b) shows a complete ten neuron qCLIF model layout with an accumulator. Further, all neurons are mapped at the right top part of the floor plan since a weight accumulator is located at the left part. Grouping the neurons in integrated circuit layouts enhances signal integrity by minimizing transmission distances and noise interference. This approach optimizes power distribution and reduces noise, while also streamlining the design and manufacturing process. This 10 qCLIF RSNN layer was subjected to simulations at various clock frequencies

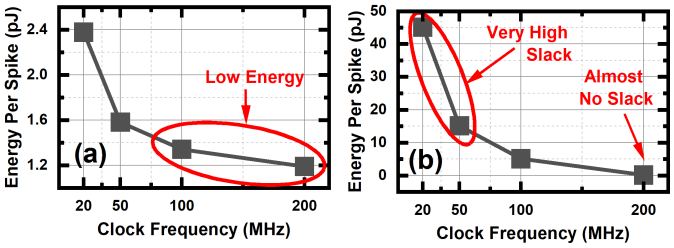

Fig. 8. Performance evaluation of a 10 qCLIF neuron layer: (a) The relationship between clock frequency and energy per spike, showing lower energy efficiency at higher frequencies. (b) Clock frequency versus timing slack, illustrating the trade-off between speed and slack available for completing the operation.

to evaluate network scalability. At a high clock frequency of 200 MHz, the network exhibited a minimal slack of 0.15 ns, albeit at the expense of increased power consumption. Conversely, a lower clock frequency of 20 MHz significantly reduced power consumption but increased the energy per spike. A median frequency of 100 MHz was found to offer a balanced trade-off, achieving a slack time of 5.10 ns and an energy per spike of 1.342 pJ as can be seen from 8.

To assess the scalability of the proposed design, it was applied to a larger network of 200 neurons, akin to a similar-sized RSNN used for DVS gesture classification, and operated at 100 MHz. With an 8-bit precision configuration, the network achieved a timing slack of 4.07 ns, occupying an area of 1.925 × 1.925 mm² and registering an energy consumption of 17.9 pJ per spike. A subsequent reduction in precision to 4 bits resulted in a smaller area of 1.365 × 1.365 mm² and lowered the energy per spike to 8.7 pJ. This demonstrates the design's adaptability to larger networks. Notably, the decrease in precision led to approximately a 50% reduction in energy per spike and, interestingly, an increased timing slack, suggesting the potential for higher clock frequencies. As shown in Table II, both area and power requirements increase sub-linearly with the number of neurons and synapses, further indicating the scalability of the design methodology.

These results suggest that the qCLIF neuron model is a viable candidate for high-speed, energy-efficient neuromorphic computing applications. Compared to previous works, as seen in Table III, while our design does not boast the highest neuron count, it introduces a more complex neuron model within a recurrent SNN framework for the first time while achieving lower energy per spike. Despite the effectiveness of this approach, it deviates from the asynchronous operation ideal in fully digital neuromorphic systems, as synchronized functioning is required for both apical and somatic compartments. The design's reliance on a digital accumulator, occupying substantial layout space, suggests potential for refinement. Space-efficient alternatives, such as sparse accumulator or memristor crossbar architectures output can be directly connected to the qCLIF digital hardware. The exploration of smaller technology nodes could yield further improvements.

TABLE II
QCLIF NEURAL MODEL COMPUTING PERFORMANCE METRICS (@ 1.1V)

No. of qCLIF	**Clock Freq** (MHz)	**Synapse Count, Precision**	**Area** (mm^2) (LXW)	**Slack (ns)**	**Switching Power** (mW)	**Internal Power** (mW)	**Leakage Power** (mW)	**Total Power** (mW)	**Energy Per Spike** (pJ)
1	100	-	0.029 X 0.030	5.62	0.020	0.041	0.016	0.077	0.773
10	20	250, 8bit	0.125 X 0.125	45.13	0.079	0.130	0.266	0.475	2.377
10	50	250, 8bit	0.125 X 0.125	15.10	0.199	0.326	0.266	0.790	1.581
10	100	250, 8bit	0.125 X 0.125	5.10	0.397	0.651	0.266	1.315	1.342
10	200	250, 8bit	0.125 X 0.125	0.15	0.805	1.306	0.268	2.380	1.190
200	100	82K, 4bit	1.365 X 1.365	6.45	72.300	70.4	31.5	174.0	8.7
200	50	82K, 8bit	1.925 X 1.925	14.05	79.500	70.8	64.0	214.0	21.4
200	100	82K, 8bit	1.925 X 1.925	4.07	153.000	141.0	63.8	358.0	17.9

TABLE III
COMPARISON OF VARIOUS SNN HARDWARE

	[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**

IV. CONCLUSION

In this study, we propose a qCLIF neuron model featuring variable precision utilizing networks sparse activity. We implemented a scalable, reconfigurable qCLIF neuron layer, marking the first hardware realization of a context-based recurrent spiking neuron layer in the digital domain. These designs are evaluated at a 45nm technology node through synthesis and layout. Our evaluations across different operating frequencies aimed to balance computational efficiency and hardware performance optimally. This work lays a step towards digital efficient neuromorphic hardware systems.

ACKNOWLEDGMENT

The authors appreciate discussions with R. Legenstein, G. H. Hutchinson, T. Bhattacharya. This study is supported by the USA National Science Foundation award #2318152.

REFERENCES

[1] Dennis V Christensen, Regina Dittmann, Bernabe Linares-Barranco, Abu Sebastian, Manuel Le Gallo, Andrea Redaelli, Stefan Slesazeck, Thomas Mikolajick, Sabina Spiga, Stephan Menzel, et al. 2022 roadmap on neuromorphic computing and engineering. *Neuromorphic Computing and Engineering*, 2(2):022501, 2022.

[2] Catherine D Schuman, Shruti R Kulkarni, Maryam Parsa, J Parker Mitchell, Prasanna Date, Bill Kay. Opportunities for neuromorphic computing algorithms and applications. *Nature Computational Science*, 2(1):10–19, 2022.

[3] Romain Ferrand, Maximilian Baronig, Thomas Limbacher, Robert Legenstein. Context-Dependent Computations in Spiking Neural Networks with Apical Modulation. In *International Conference on Artificial Neural Networks*, pages 381–392, 2023.

[4] Sumit Bam Shrestha, Garrick Orchard. SLAYER: Spike Layer Error Reassignment in Time. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[5] Weihua He, YuJie Wu, Lei Deng, Guoqi Li, Haoyu Wang, Yang Tian, Wei Ding, Wenhui Wang, Yuan Xie. Comparing SNNs and RNNs on neuromorphic vision datasets: Similarities and differences. *Neural Networks*, 132:108-120, 2020.

[6] Xueyuan She, Saurabh Dash, Saibal Mukhopadhyay. Sequence Approximation using Feedforward Spiking Neural Network for Spatiotemporal Learning: Theory and Optimization Methods. In *International Conference on Learning Representations*, 2022.

[7] Ali Samadzadeh, Fatemeh Sadat Tabatabaei Far, Ali Javadi, Ahmad Nickabadi, Morteza Haghir Chehreghani. Convolutional spiking neural networks for spatio-temporal feature extraction. *Neural Processing Letters*, pages 1–17, 2023.

[8] Anand Subramoney, Khaleelulla Khan Nazeer, Mark Schöne, Christian Mayr, David Kappel. Efficient recurrent architectures through activity sparsity and sparse back-propagation through time. In *The Eleventh International Conference on Learning Representations*, 2023.

[9] Mingqing Xiao, Qingyan Meng, Zongpeng Zhang, Di He, Zhouchen Lin. Online Training Through Time for Spiking Neural Networks. In *Advances in Neural Information Processing Systems*, volume 35, pages

20717–20730, 2022.
[10] A. Shaban, S. S. Bezugam, and M. Suri, "An adaptive threshold neuron for recurrent spiking neural networks with nanodevice hardware implementation," *Nature Communications*, vol. 12, no. 1, p. 4234, 2021.
[11] G. Bellec, D. Salaj, A. Subramoney, R. Legenstein, and W. Maass, "Long short-term memory and learning-to-learn in networks of spiking neurons," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.
[12] G. Bellec, F. Scherr, A. Subramoney, et al., "A solution to the learning dilemma for recurrent networks of spiking neurons," *Nature Communications*, vol. 11, p. 3625, 2020.
[13] B. Yin, F. Corradi, and S. M. Bohté, "Accurate online training of dynamical spiking neural networks through Forward Propagation Through Time," *Nature Machine Intelligence*, vol. 5, pp. 518–527, 2023. [Online]. Available: https://doi.org/10.1038/s42256-023-00650-4
[14] M. Baronig and R. Legenstein, "Context association in pyramidal neurons through local synaptic plasticity in apical dendrites," *Frontiers in Neuroscience*, vol. 17, p. 1276706, 2023.
[15] J. E. Stine, I. Castellanos, M. Wood, J. Henson, F. Love, W. R. Davis, P. D. Franzon, M. Bucher, S. Basavarajaiah, J. Oh et al. Freepdk: An open source variation-aware design kit. In *2007 IEEE International Conference on Microelectronic Systems Education (MSE'07)*, pages 173–174, IEEE, 2007.
[16] A. Amir, B. Taba, D. Berg, T. Melano, J. McKinstry, C. Di Nolfo, T. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza et al. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7243–7252, 2017.
[17] B. Cramer, Y. Stradmann, J. Schemmel, and F. Zenke. The heidelberg spiking data sets for the systematic evaluation of spiking neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2744–2757, 2020.
[18] Pavan Kumar Chundi, et al. Always-on sub-microwatt spiking neural network based on spike-driven clock-and power-gating for an ultra-low-power intelligent device. *Frontiers in Neuroscience*, 15:684113, 2021.
[19] M. Koo, G. Srinivasan, Y. Shim, and K. Roy. sBSNN: stochastic-bits enabled binary spiking neural network with on-chip learning for energy efficient neuromorphic computing at the edge. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 67:2546–2555, 2020. doi: 10.1109/TCSI.2020.2979826.
[20] J. Park, Juyun L., and Dongsuk J. 7.6 A 65nm 236.5nJ/classification neuromorphic processor with 7.5% energy overhead on-chip learning using direct spike-only feedback. In *Proceedings of the 2019 IEEE International Solid-State Circuits Conference - (ISSCC)*, pages 140–142, San Francisco, CA, 2019. doi: 10.1109/isscc.2019.8662398.
[21] G. K. Chen, K. Raghavan, H. Ekin Sumbul, C. K. Phil, and K. K. Ram. A 4096-Neuron 1M-Synapse 3.8PJ/SOP spiking neural network with on-chip STDP learning and sparse weights in 10NM FinFET CMOS. In *IEEE Symposium on VLSI Circuits, Honolulu, HI*, pages 255–256, 2018. doi: 10.1109/VLSIC.2018.8502423.
[22] F. Akopyan, S. Jun, C. Andrew, A. I. Rodrigo, A. John, M. Paul, et al. TrueNorth: design and tool flow of a 65 mW 1 Million neuron programmable neurosynaptic chip. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 34:1537–1557, 2015. doi: 10.1109/tcad.2015.2474396.
[23] C. Ganguly, S. Bezugam, E. Abs, M. Payvand, S. Dey, and M. Suri. Spike frequency adaptation: bridging neural models and neuromorphic applications. *Communications Engineering*, 3(1):22, 2024. DOI: 10.1038/s44172-024-00165-9.