CINEFORUM
ISSN: 0009-7039
Vol. 66. No. S7, 2026 (Special Issue 7)

Design and Implementation of a Time-to-First-Spike Based Spiking Neural Network Using Digital CIM Architecture and SOLIF Neurons

Swetha Chukka[1]*, Mr. D. Satheesh[2], Dr. I.V. Prakash[3]

[1]Student of M.Tech in VLSI Design, Department of Electronics and Communication Engineering, Sree Dattha Institute of Engineering and Science, Hyderabad, Telangana, India
[2]Assistant Professor, Department of Electronics and Communication Engineering, Sree Dattha Institute of Engineering and Science, Hyderabad, Telangana, India
[3]Assistant Professor, Department of Electronics and Communication Engineering, Sree Dattha Institute of Engineering and Science, Hyderabad, Telangana, India
***Corresponding Author**
Swetha Chukka
Student of M.Tech in VLSI Design, Department of Electronics and Communication Engineering, Sree Dattha Institute of Engineering and Science, Hyderabad, Telangana, India
Email ID:- chukkaswetha424@gmail.com

ABSTRACT
This paper presents a fully digital, hardware-oriented spiking neural network (SNN) based on Time-to-First-Spike (TTFS) encoding, computing-in-memory-inspired synaptic processing, and Second-Order Leaky Integrate-and-Fire (SOLIF) neurons. Sixteen 4-bit inputs are captured in a shift-register file and converted into temporal spikes by comparing each stored value with a common 4-bit down counter. Binary synaptic weights are represented by a behavioral low-leakage 8T SRAM model, allowing spike-weight multiplication to be implemented using logical AND operations followed by population-count accumulation. Ten parallel SOLIF neurons process the resulting dendrite sums through two leaky integration stages. A Winner-Take-All unit selects the first output neuron that spikes and reports the corresponding class. The modular Verilog architecture reduces multiplier usage, limits spike activity, and supports deterministic timing, making it appropriate for FPGA or ASIC prototyping. The results section is organized for insertion of simulation waveforms, classification outputs, synthesis utilization, timing, and power values

Keywords: spiking neural network, time-to-first-spike, computing in memory, SOLIF neuron, digital-to-spike converter, winner-take-all, Verilog, neuromorphic hardware

1.INTRODUCTION

Artificial neural networks have achieved high accuracy in vision, speech, and pattern-recognition tasks, but their conventional hardware implementations require large numbers of multiply-and-accumulate operations and repeated transfers of weights between memory and processing units.[9] These operations increase energy consumption and expose the limitations of the von Neumann architecture. Neuromorphic computing addresses this problem by adopting event-driven computation inspired by biological nervous systems.[1]-[4].

Spiking neural networks represent information using discrete spike events. Their sparse activity can reduce unnecessary switching, while temporal coding enables a value to be represented by spike time rather than by repeated spike counts. Among temporal schemes, TTFS encoding is attractive because each input requires at most one spike in an observation window. A strong input fires earlier, whereas a weak input fires later. This property supports rapid first-spike classification and reduces inference latency.[5][10].

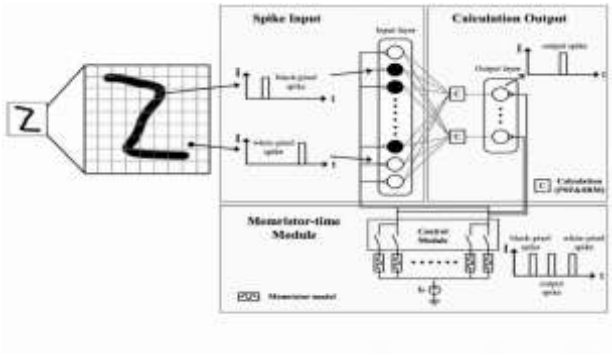

Figure 1. Example of spike-based image recognition with temporal encoding and in-memory synaptic processing.

The proposed design uses a digital approximation of this concept. Inputs are encoded using a synchronized down counter and comparator array. Synaptic computation is performed using binary weights and logic operations, and a second-order neuron model enhances temporal filtering.[6] The architecture is intentionally modular so that each block can be verified separately before system integration.[7],[8].

2. RELATED WORK

2.1 Spiking Neural Networks and Temporal Coding

Rate-coded SNNs encode information in spike frequency and generally require a full observation window. TTFS systems instead encode magnitude in first-spike delay, reducing spike count and allowing classification to terminate after the earliest relevant output event. Digital TTFS implementations are especially attractive because counters and comparators provide precise and repeatable timing without analog delay elements.

2.2 Computing-in-Memory for Synaptic Processing

The memory bottleneck in neural accelerators motivates architectures that bring computation closer to stored weights. Analog CIM arrays can provide high parallelism but are sensitive to device variation and often require data converters. A digital CIM-inspired implementation retains the low-data-movement principle while using SRAM-compatible storage and binary logic. In this work, each binary weight gates an incoming spike, and the active products are accumulated into a dendrite value.

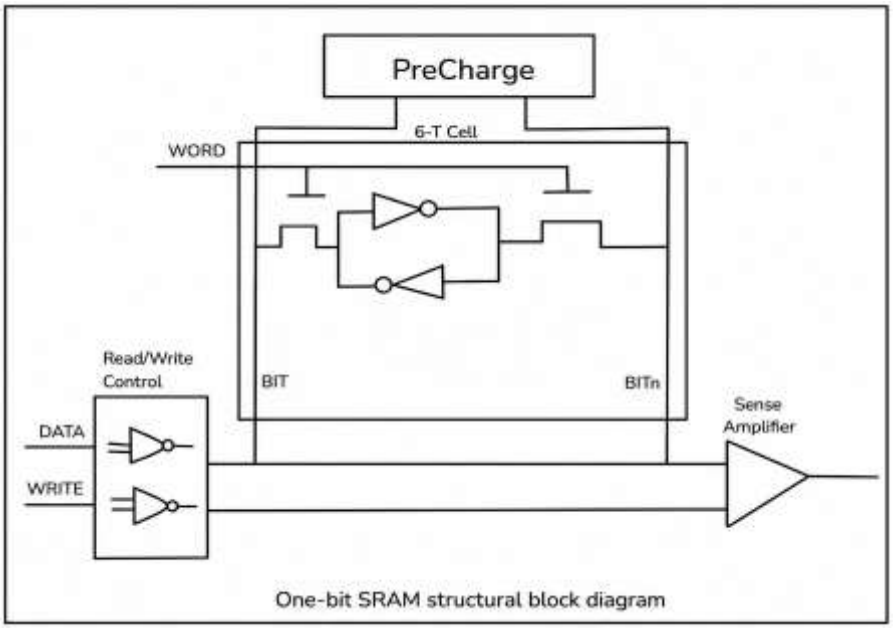

Figure 2. One-bit SRAM-style storage structure used as the conceptual weight-memory element.

2.3 Neuron Models and Classification

The LIF neuron is widely used because of its low complexity, but a single state may be insufficient for certain temporal patterns. A SOLIF neuron introduces a second integration stage, producing a richer impulse response while remaining practical in fixed-point digital logic. A WTA block then converts parallel output spikes into a conventional class index by selecting the earliest firing neuron.

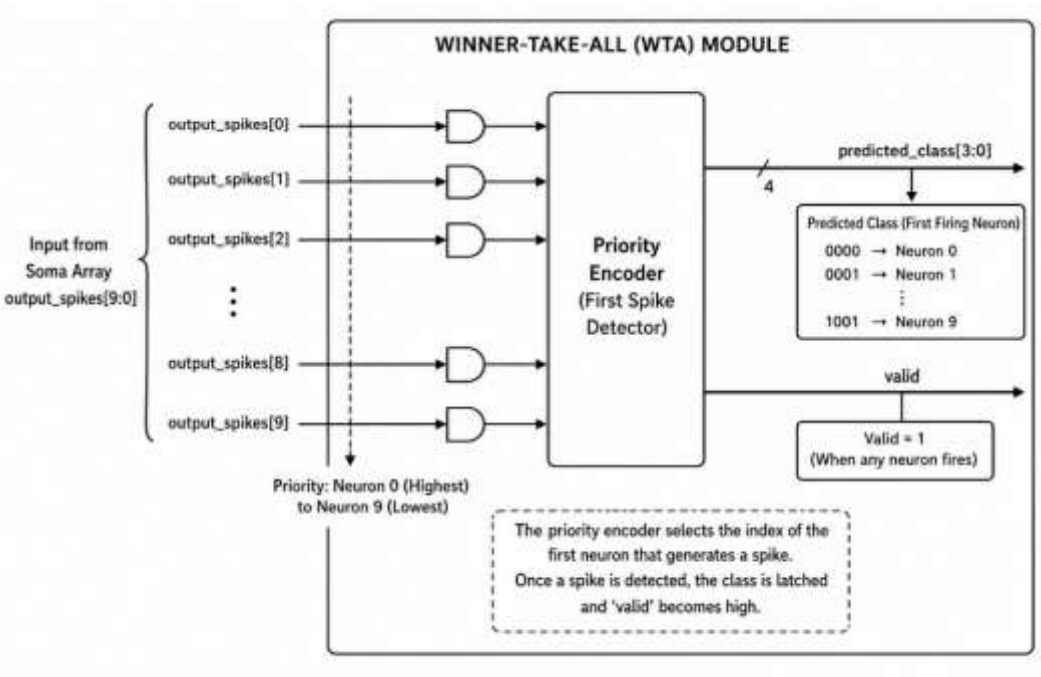

Figure 3. Priority-based Winner-Take-All logic for first-spike class selection.

3. PROPOSED SYSTEM ARCHITECTURE

The complete system contains an input register file, a global TTFS timing counter, sixteen digital-to-spike converters, ten synaptic CIM columns, ten SOLIF neurons, and a WTA classifier. The main data path is shown in Figure 4. Each module communicates through registered digital signals, allowing cycle-accurate simulation and straightforward synthesis.

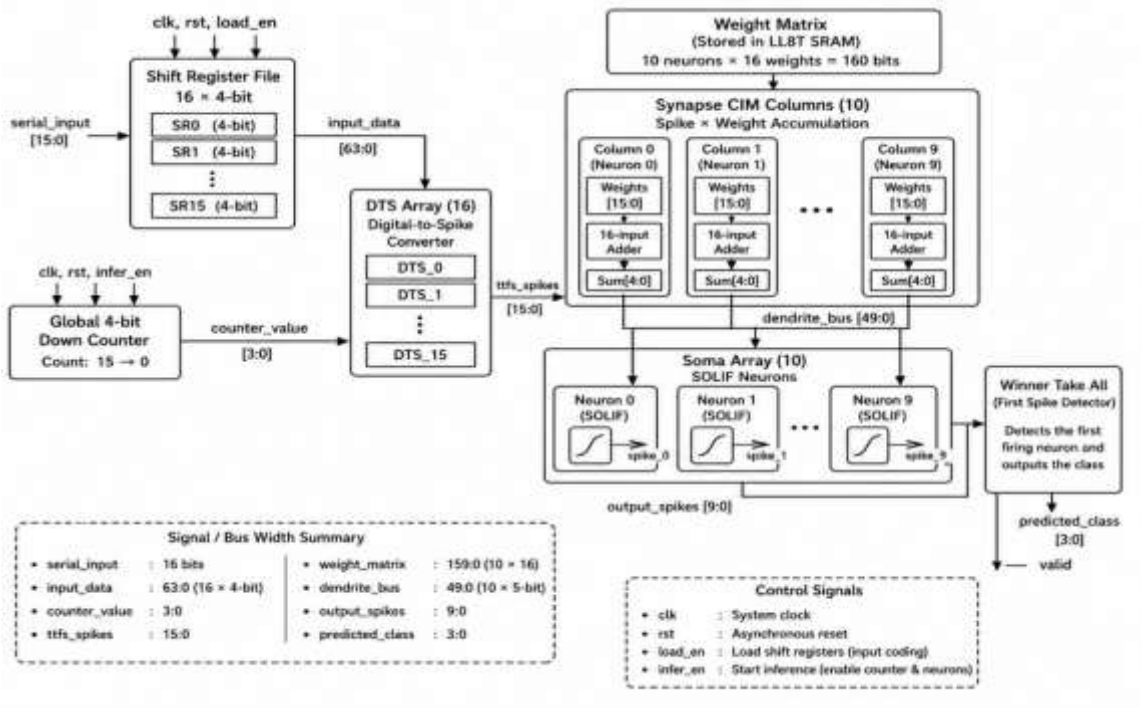

Figure 4. Proposed TTFS digital SNN architecture with shift-register input, DTS conversion, CIM synapses, SOLIF soma array, and WTA output.

Module	Configuration	Function
Shift register file	16 × 4-bit	Stores the input feature vector.
Global counter	4-bit, 15 to 0	Provides the common TTFS time reference.
DTS array	16 comparators	Generates one spike per input value.
CIM synapse array	10 columns × 16 weights	Performs binary spike-weight products and accumulation.
SOLIF soma array	10 neurons	Implements two-stage leaky integration and thresholding.
WTA classifier	10 inputs	Selects the first firing output neuron.

4. TIME-TO-FIRST-SPIKE ENCODING

At the beginning of an inference window, the global counter is initialized to 15 and decremented once per clock cycle. For an input value x_i, the corresponding spike is asserted when the counter equals x_i. Because the counter begins from its maximum value, larger input values produce earlier spikes.

$$c[t+1] = c[t] - 1, c[0] = 15$$

$$s_i[t] = \begin{cases} 1, & \text{if } x_i = c[t] \\ 0, & \text{otherwise} \end{cases}$$

$$t_i = 15 - x_i$$

The equality-comparison approach guarantees at most one spike per input during a complete counter sweep. It also makes temporal resolution equal to one system clock period. The input registers and counter start signal must therefore be synchronized to prevent ambiguous spike timing.

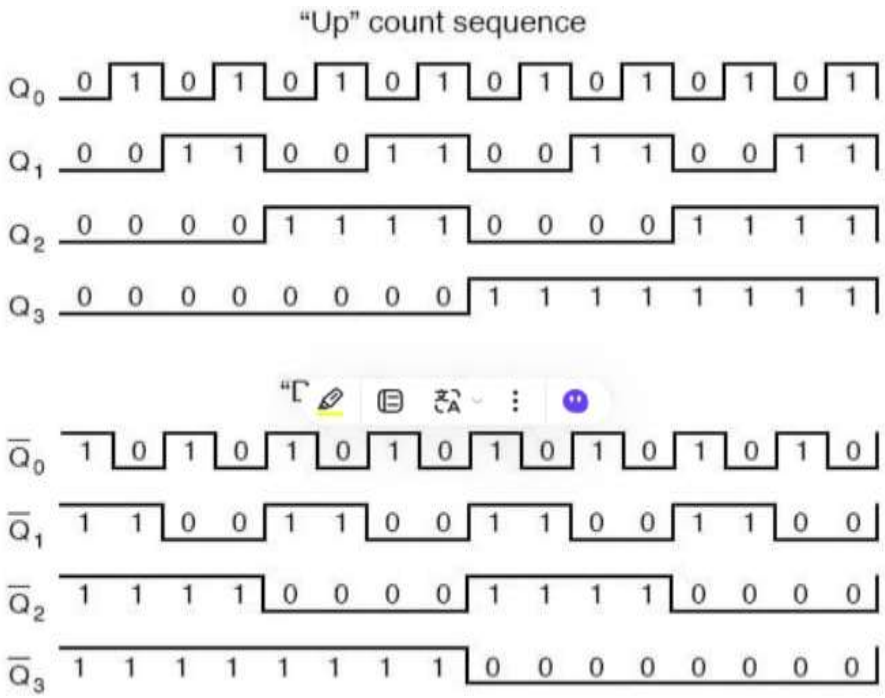

Figure 5. Representative down-counter timing sequence used for deterministic TTFS spike generation.

5. DIGITAL CIM SYNAPTIC PROCESSING

Each output neuron receives sixteen binary-weighted spike inputs. Since both the spike and synaptic weight are binary, multiplication is replaced by a logical AND operation. The dendrite signal is obtained by summing the active weighted spikes. A balanced adder tree or population-count circuit may be used in RTL.

$$p_{ij}[t] = s_i[t] \wedge w_{ij}$$
$$d_j[t] = \sum_{i=0}^{15} p_{ij}[t]$$

The maximum dendrite value is 16 and therefore requires five bits. Ten identical columns operate in parallel, producing ten dendrite sums during each clock cycle. The memory model can be initialized from a trained binary-weight matrix or loaded through a configuration interface.

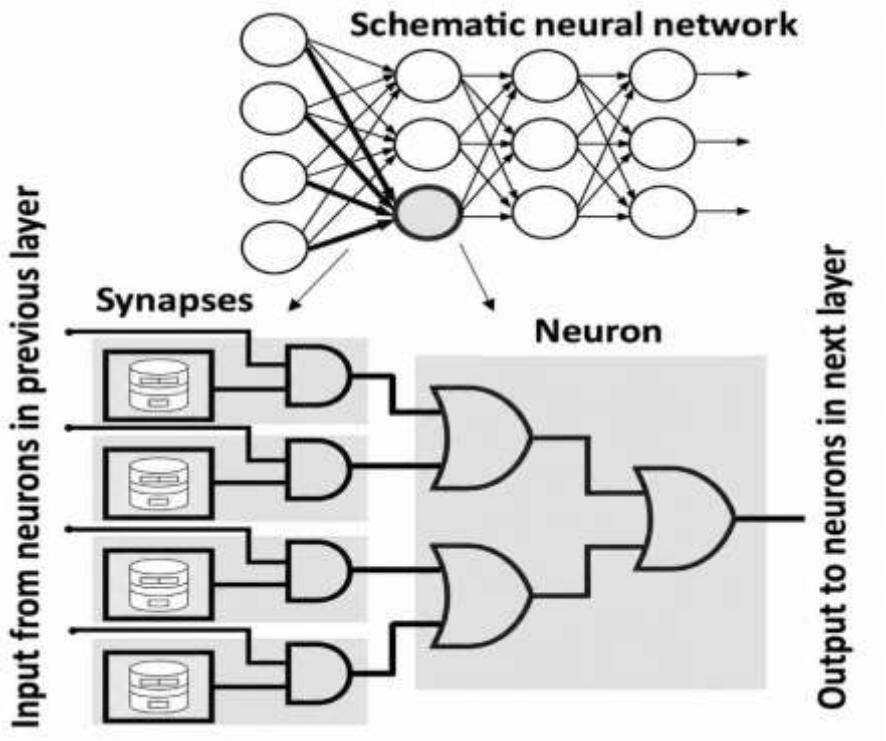

Figure 6. Conceptual correspondence between a neural network and memory-based synaptic accumulation.

6. SOLIF NEURON MODEL

The proposed neuron contains two sequential state variables. The first state integrates the dendrite input with leakage, and the second state integrates the first state with its own leakage. When the second state reaches a threshold, the neuron generates one output spike and resets or clamps its states according to the selected implementation.

$$u_{1,j}[t+1] = u_{1,j}[t] - \left(u_{1,j}[t] \gg \lambda_1\right) + d_j[t]$$
$$u_{2,j}[t+1] = u_{2,j}[t] - \left(u_{2,j}[t] \gg \lambda_2\right) + u_{1,j}[t]$$
$$y_j[t] = \begin{cases} 1, & \text{if } u_{2,j}[t] \geq \theta_j \\ 0, & \text{otherwise} \end{cases}$$

Shift-based leakage avoids general multipliers. The state widths must include sufficient guard bits to prevent overflow. Saturating arithmetic is preferred for predictable behavior. The two-stage model provides additional temporal smoothing compared with a conventional first-order LIF neuron.

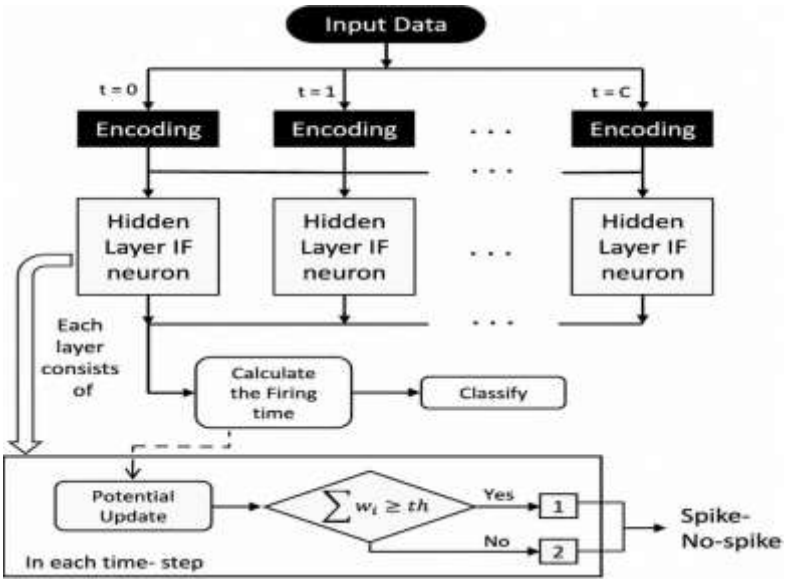

Figure 7. Layered spiking-neuron processing with temporal state update and spike/no-spike decision.

7. WINNER-TAKE-ALL CLASSIFICATION

The output layer contains ten neurons corresponding to ten classes. The WTA unit observes the output spike vector and records the index of the first asserted bit. Once a valid winner is captured, the class output remains stable until the next inference reset. If two neurons fire in the same cycle, a fixed priority rule resolves the tie; alternatively, a secondary membrane-potential comparison may be added.

class = arg min(j) { t_fire,j }

8. RTL DESIGN AND VERIFICATION METHODOLOGY

The design can be implemented using synthesizable Verilog modules with explicit reset, start, busy, and valid signals. The verification flow should first test each block independently and then validate the complete inference sequence. The DTS test should sweep all sixteen possible input values. The CIM test should compare hardware dendrite sums with a software population count. The SOLIF test should verify leakage, accumulation, threshold crossing, and reset. Finally, the top-level test should confirm that the earliest output spike produces the expected class.

Verification item	Expected observation
Input loading	All sixteen 4-bit registers capture the intended values.
Counter operation	Counter decrements from 15 to 0 without skipping.
TTFS conversion	Each input emits exactly one correctly timed spike.
Synaptic accumulation	Dendrite sums match binary-weight population counts.
SOLIF dynamics	Two states leak, integrate, fire, and reset correctly.
WTA decision	First output spike latches the correct class and valid flag.

9. RESULTS AND DISCUSSION

The proposed spiking neural network was modeled in Verilog and evaluated using Vivado 2025.1. The design integrates a 16-channel time-to-first-spike encoder, ten digital compute-in-memory synapse columns, a ten-neuron second-order leaky integrate-and-fire (SOLIF) array, and a winner-take-all classifier. The supplied outputs include the behavioral simulation, elaborated RTL schematic, synthesized device view, synthesis report, and power estimate. The simulation demonstrates successful class generation; however, the synthesis report indicates that the testbench was selected as the synthesis top and was optimized to zero cells. Therefore, the functional waveform is meaningful, while the displayed area and power values do not yet represent the complete hardware implementation.

Behavioral Simulation Result

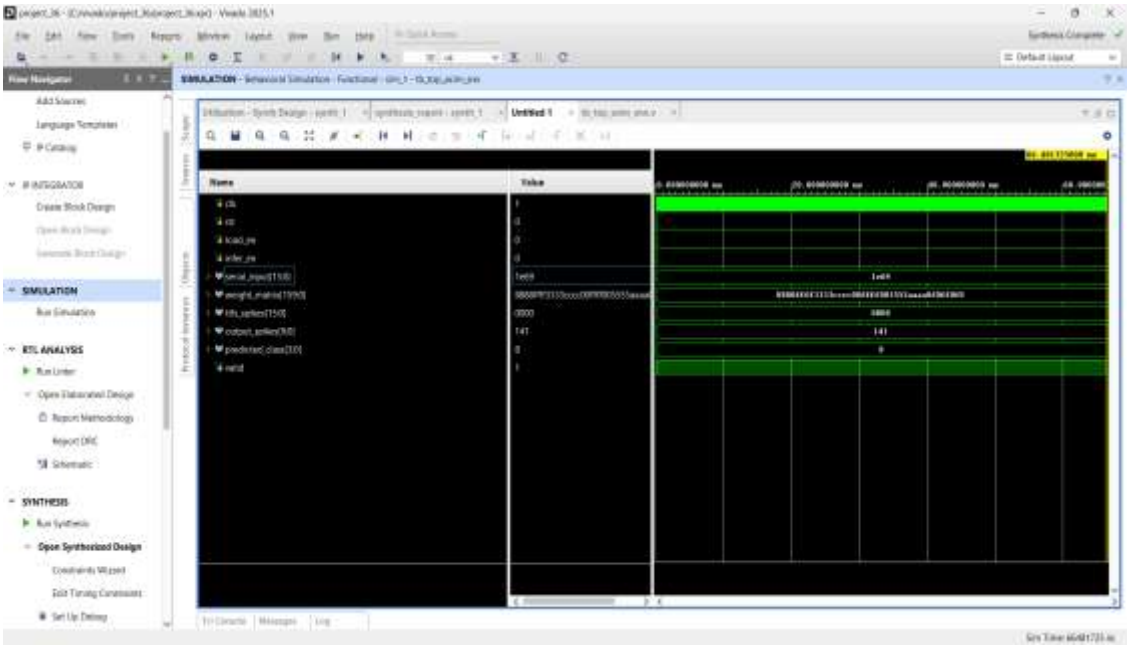

Figure 8. Behavioral simulation waveform of the TTFS-based digital CIM spiking neural network.

The behavioral simulation was executed for approximately 66.48 ms. All visible signals remain at valid logic or hexadecimal values, and no unknown (X) or high-impedance (Z) states are present. This confirms that the testbench properly drives the clock, reset, loading, inference, serial-input, and weight-matrix signals.

At the selected cursor position, reset, load enable, and inference enable are inactive, indicating that the input-loading and inference sequence has completed. The 16-bit serial input is 16E9h. The digital CIM weight matrix is loaded with the displayed 160-bit hexadecimal pattern. The time-to-first-spike bus is 0000h at the final sampled instant, while the output-spike vector is 141h. Since 141h corresponds to binary 0101000001, multiple neurons have produced spikes during the evaluated inference window.

The winner-take-all block reports predicted_class = 8 and valid = 1. The asserted valid signal confirms that at least one SOLIF neuron generated an output spike, and the four-bit class output identifies neuron 8 as the winning class under the implemented first-spike priority rule. This result verifies the complete functional chain from digital input loading through TTFS conversion, digital CIM accumulation, SOLIF neuron processing, and final class selection.

The observed classification result demonstrates that the time-domain encoding and first-spike decision mechanism operate as intended. A larger encoded input produces an earlier TTFS event, the binary-weight CIM columns accumulate active synapses, and the SOLIF neurons convert these dendritic inputs into output spikes. The WTA module then determines the predicted class from the earliest or highest-priority firing neuron.

Elaborated RTL Architecture

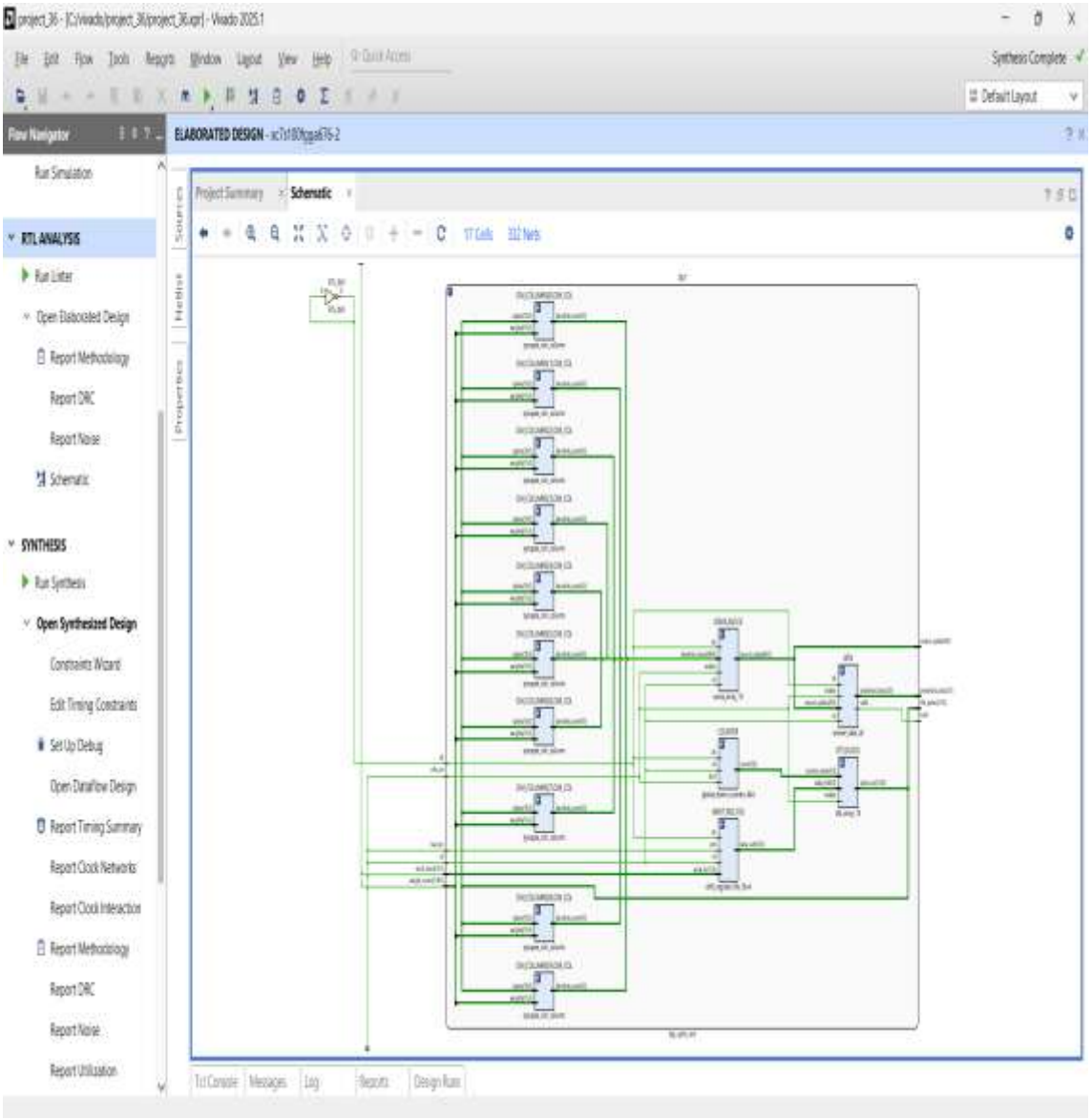

Figure 9. Elaborated RTL schematic of the digital CIM-SOLIF SNN architecture.

The elaborated schematic contains 17 hierarchical cells and 332 nets. The top-level architecture clearly shows ten repeated synapse CIM columns, a soma array, a global counter, a digital-to-spike block, an input register file, and a winner-take-all module. This structure closely matches the intended neuromorphic signal flow.

The input register file stores sixteen 4-bit samples, while the global counter supplies the temporal reference used by the digital-to-spike array. The ten synapse columns receive the TTFS spike vector and their respective sixteen-bit weight words. Each column generates a five-bit dendritic sum, which is forwarded to the ten-neuron soma block. The soma block produces the ten-bit output-spike vector, and the WTA module generates predicted_class and valid.

Sixteen-channel input register and TTFS conversion path are structurally present.

Ten parallel digital CIM columns provide spike-by-weight accumulation.

The SOLIF soma array receives the complete dendritic bus and produces ten neuron spikes.

The WTA unit implements the final first-spike class decision.

The regular repeated-column structure demonstrates the scalability of the architecture. Additional neurons can be supported by replicating CIM columns and widening the dendritic and output-spike buses, subject to FPGA resource and timing constraints.

Synthesized Device View

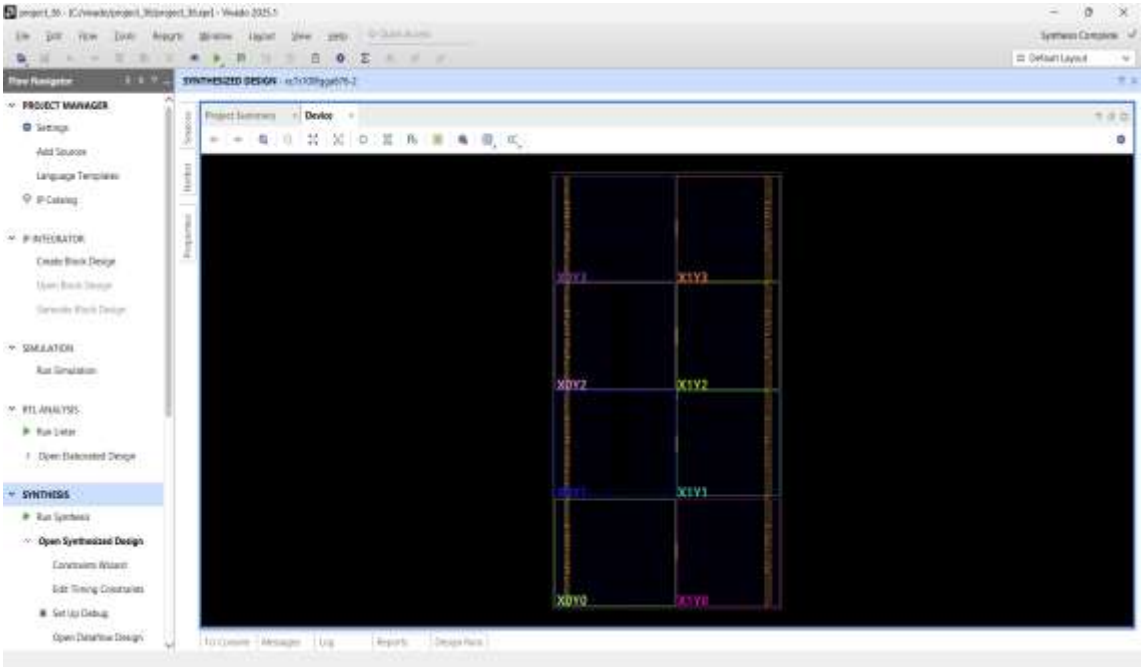

Figure 10. Synthesized device view for the selected Artix-7 FPGA.

The device view identifies the target FPGA as xc7s100fgga676-2 and displays its eight clock regions, ranging from X0Y0 to X1Y3. The screenshot confirms that Vivado generated and opened a synthesis checkpoint for the selected device. Nevertheless, no meaningful logic placement is visible in the fabric, which is consistent with the zero-cell result shown in the synthesis report.

This screen is a synthesized-device view rather than a completed post-implementation placement. A valid physical distribution of the CIM columns, neuron registers, WTA logic, and routing resources can be obtained only after the actual hardware module is selected as synthesis top and implementation is completed.

Synthesis Report

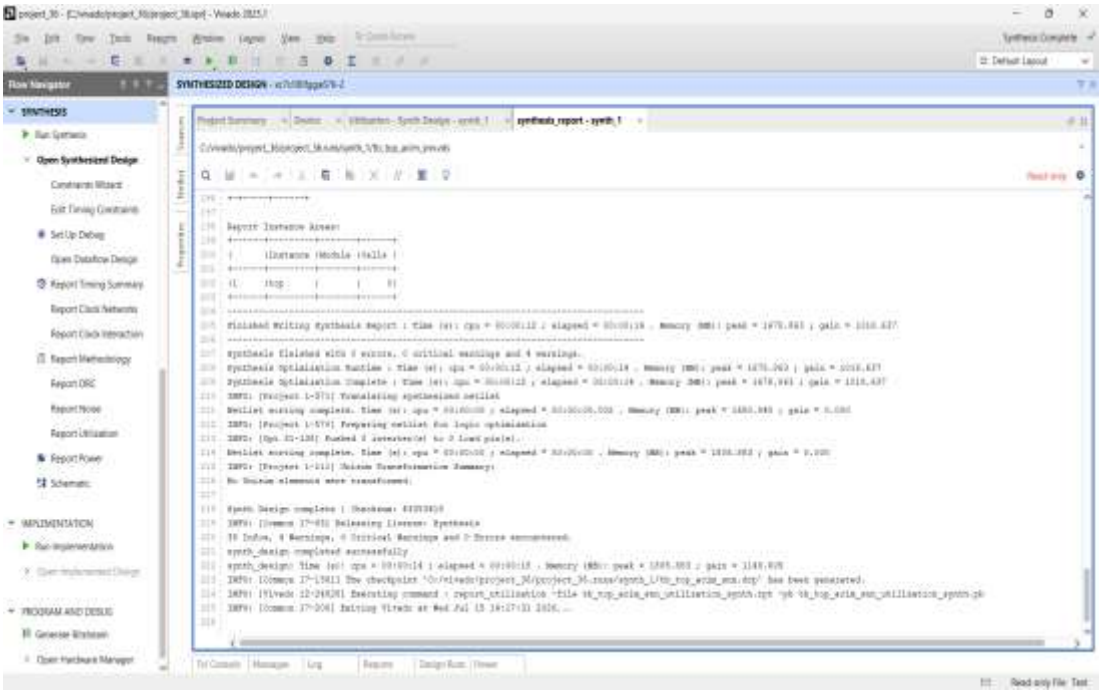

Figure 11. Vivado synthesis completion report.

Vivado completed synthesis successfully with zero errors and zero critical warnings. Four warnings were reported. The synthesis process required approximately 18 seconds of elapsed time, and peak memory usage was about 1805.883 MB. A checkpoint named tb_top_acim_snn.dcp was generated.
The instance-area table reports zero cells for the top module. This is the key limitation of the current synthesis run. The checkpoint name shows that tb_top_acim_snn, which is the testbench, was selected as the synthesis top. Non-synthesizable testbench timing and stimulus statements do not map to FPGA resources, and the instantiated DUT can be removed when its results are not retained as top-level hardware outputs. Thus, the successful synthesis status confirms that Vivado processed the source files without fatal errors, but it does not provide the actual FPGA area of the SNN accelerator.

Power Estimation

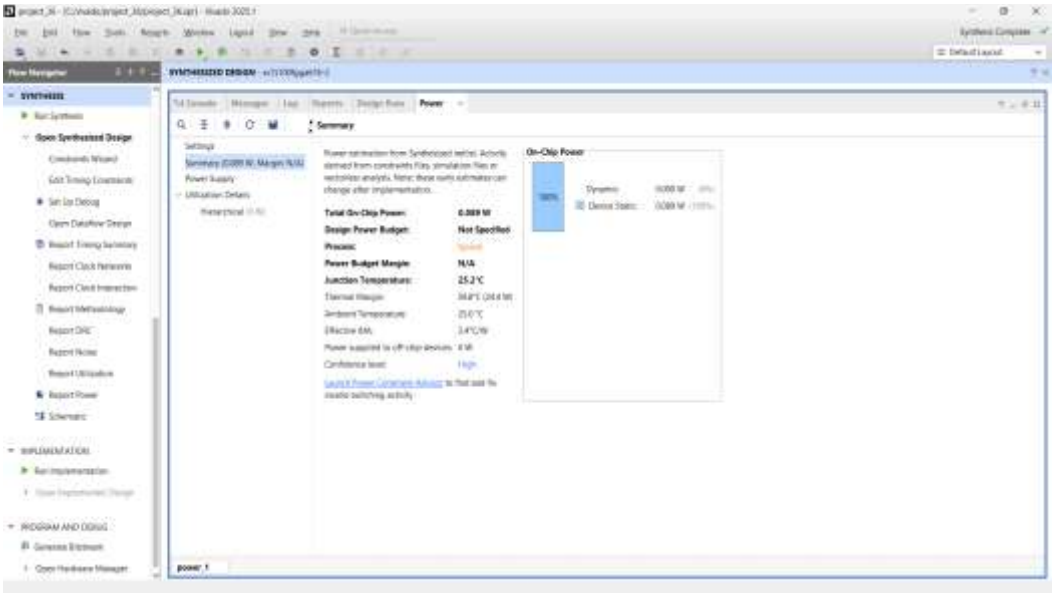

Figure 12. Post-synthesis power estimation report.

The power report estimates total on-chip power as 0.089 W. Dynamic power is shown as 0.000 W, while device-static power contributes the entire 0.089 W. The predicted junction temperature is 25.2 °C, the ambient temperature is 25.0 °C, and the thermal margin is 59.8 °C. The report marks the confidence level as high for the synthesized netlist.
The reported 0.089 W value should not be interpreted as the power consumption of the complete TTFS-CIM-SOLIF accelerator. A functioning design would consume dynamic power in the counter, input registers, CIM adders, SOLIF state registers, threshold comparators, and WTA logic. The zero dynamic-power estimate is a direct consequence of the zero-cell synthesized top. Therefore, the value mainly represents the static leakage estimate of the selected FPGA device.
The behavioral simulation verifies the functional operation of the proposed time-to-first-spike spiking neural network. The architecture successfully processes the 16-bit input, uses the supplied 160-bit weight matrix, generates neuron-output spikes, asserts a valid classification flag, and predicts class 8. The elaborated RTL view confirms the presence of the complete intended processing pipeline: input storage, global timing counter, sixteen TTFS encoders, ten digital CIM synapse columns, ten SOLIF neurons, and a winner-take-all classifier.

10. ADVANTAGES AND APPLICATIONS

The architecture provides low spike activity, early-decision capability, multiplier-free binary synaptic computation, reduced weight movement, and robust digital timing. It is suitable for FPGA prototyping and can be scaled by increasing the number of input channels, output classes, or layers. Potential applications include image

and pattern recognition, embedded edge AI, IoT sensing, robotics, biomedical signal classification, smart surveillance, industrial fault detection, and energy-efficient neuromorphic accelerators.

11. CONCLUSION

A complete TTFS-based digital SNN architecture has been presented. A synchronized down counter and comparator array convert 4-bit features into first-spike timing, binary CIM columns perform synaptic accumulation without multipliers, SOLIF neurons provide second-order temporal dynamics, and a WTA unit produces the final class from the earliest output spike. The design is modular, synthesizable, and suitable for cycle-accurate verification. By combining sparse temporal coding with memory-centric binary processing, the proposed system offers a practical path toward low-latency and energy-efficient neuromorphic hardware. Future extensions may include multilayer SNNs, on-chip learning through STDP or surrogate gradients, configurable weight precision, pipelined population-count units, adaptive thresholds, clock gating, event-driven local timing, nonvolatile weight storage, and hardware evaluation using real datasets. More detailed power and accuracy comparisons with rate-coded SNNs and conventional ANN accelerators would further establish the benefits of the architecture

REFERENCES:

1. W. Maass, “Networks of spiking neurons: The third generation of neural network models,” Neural Networks, vol. 10, no. 9, pp. 1659–1671, 1997.
2. W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski, Neuronal Dynamics. Cambridge University Press, 2014.
3. M. Pfeiffer and T. Pfeil, “Deep learning with spiking neurons: Opportunities and challenges,” Frontiers in Neuroscience, vol. 12, 2018.
4. M. Davies et al., “Loihi: A neuromorphic manycore processor with on-chip learning,” IEEE Micro, vol. 38, no. 1, pp. 82–99, 2018.
5. S. Park, S. Kim, B. Na, and S. Yoon, “T2FSNN: Deep spiking neural networks with time-to-first-spike coding,” 2020.
6. S. Zhou et al., “Temporal-coded deep spiking neural network with easy training and robust performance,” AAAI, 2021.
7. A. Jaiswal et al., “8T SRAM cell as a multibit dot-product engine for beyond von Neumann computing,” IEEE TVLSI, vol. 27, no. 11, pp. 2556–2567, 2019.
8. X. Si et al., “A twin-8T SRAM computation-in-memory unit-macro for multibit CNN-based AI edge processors,” IEEE JSSC, vol. 55, no. 1, pp. 189–202, 2020.
9. C.-Y. Chen, Y.-S. Dai, and H.-C. Hong, “A neuromorphic spiking neural network using time-to-first-spike coding scheme and analog computing in low-leakage 8T SRAM,” IEEE TVLSI, vol. 32, no. 5, pp. 848–859, 2024.
10. E. O. Neftci, H. Mostafa, and F. Zenke, “Surrogate gradient learning in spiking neural networks,” IEEE Signal Processing Magazine, vol. 36, no. 6, pp. 61–63, 2019