Darwin: A Neuromorphic Hardware Co-Processor for Spiking Neural Networks

SHEN JunCheng[1,3], MA De[2], GU ZongHua[1*], ZHANG Ming[1], ZHU XiaoLei[3], XU XiaoQiang[1]
XU Qi[1], SHEN YangJing[2]& PAN Gang[1]
[1]*College of Computer Science, Zhejiang University, Hangzhou, 310027, China*
[2]*Key Laboratory of RF Circuits and Systems, Ministry of Education,*
Hangzhou Dianzi University, 310018, China
[3]*Institute of VLSI design, Zhejiang University, Hangzhou, 310027, China*

SCIENCE CHINA PRESS

The Leaky Integrate and Fire (LIF) Model

- A simplified model of biological neuron, where the membrane voltage V is described as:

$$C_m \frac{dV}{dt} = g_l(V_{rest} - \mathrm{V}) + \mathrm{I}$$

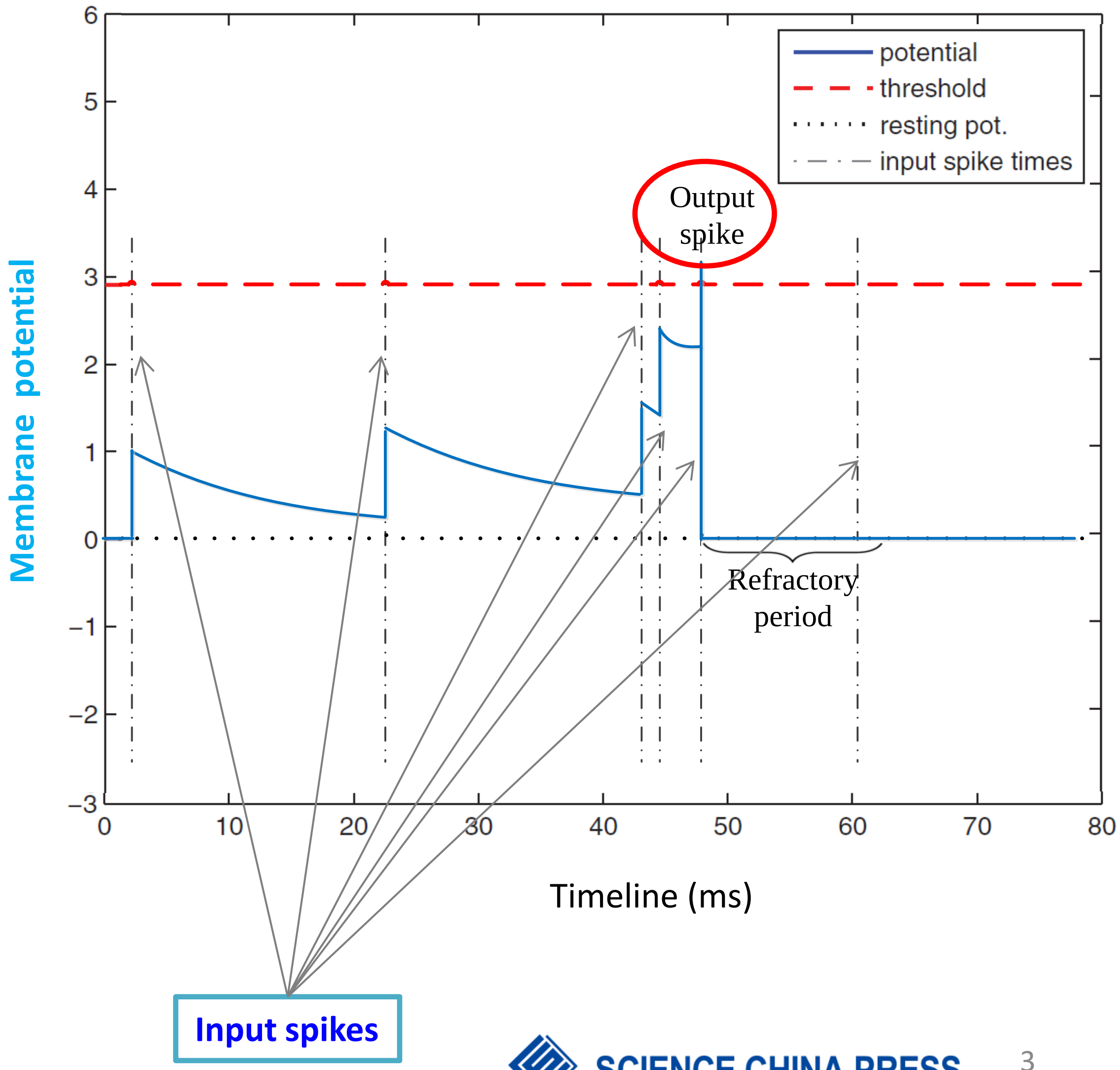
  - V_{rest} is the resting membrane potential;
  - $C_m$is the membrane capacitance;
  - g_l is the membrane conductance;
  - I is the input current.

When V rises up to reach the firing threshold V_{th}, a spike is triggered, and V rapidly rises to a large value, then reset to $V = V_{reset}$. Afterwards, there is a refractory period T_{ref}, when the neuron is not responsible to input spikes.

LIF Model: An Example

- **Membrane potential** rises upon each input spike, and gradually leaks and returns to resting membrane potential

- When **input spikes** arrive in close timing proximity, the membrane potential reaches threshold and fires an output spike



LIF Model： Discrete Time Version

- Discrete-time version is necessary to implement the LIF model in digital logic.
- Consider a post-synaptic neuron with index *j*, connected to possibly multiple pre-synaptic neurons with indices denoted as *i*. The membrane potential of neuron *j*satisfies the following discrete time equation:

$$V_j(t) \leftarrow V_j(t-1)(1-\Delta t/\tau_m) + \sum_i S_{ij} V_{max} w_{ij}$$

$$V_j(t) \leftarrow H\big(V_{th} - V_j(t)\big) \cdot V_j(t)$$

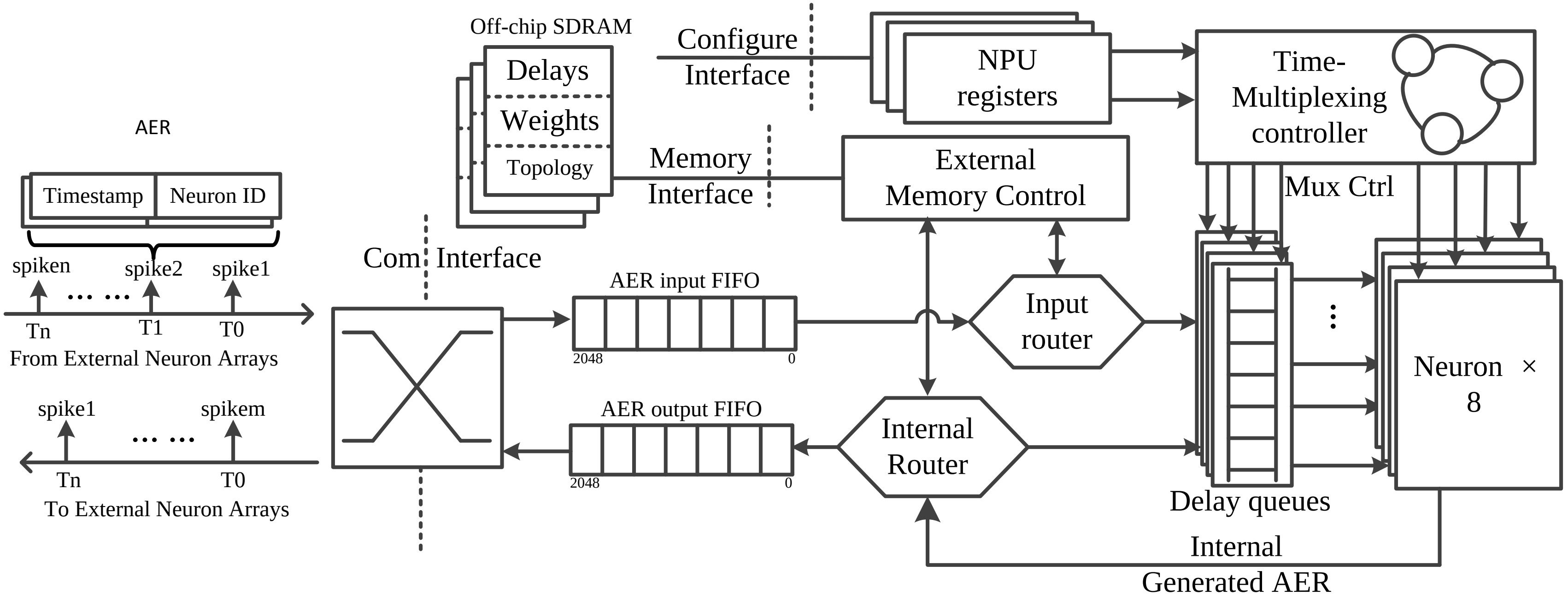
$$S_i(t) \leftarrow H(V_i(t) - V_{th})$$

  - $V_j(t)$ is the membrane potential of neuron *j* at time step *t*,
  - $\Delta t$is simulation time step size;
  - τ_m=$C_m/g_l$is time constant of the RC circuit model of the cell membrane;
  - $S_{ij} = \{0, 1\}$denotes whether neuron *i* fires a spike at time step *t*;
  - V_{max}denotes the maximum voltage change to a neuron caused by receiving an incoming spike;
  - w_{ij} indicates the weight of the synapse that connects pre-synaptic neuron *i*to the post-synaptic neuron *j*; it is positive if the synapse excitatory; negative if itis inhibitory;
  - V_{th} is the firing threshold;
  - $H(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases}$ is the unit step function.

Overall Micro-Architecture of Darwin

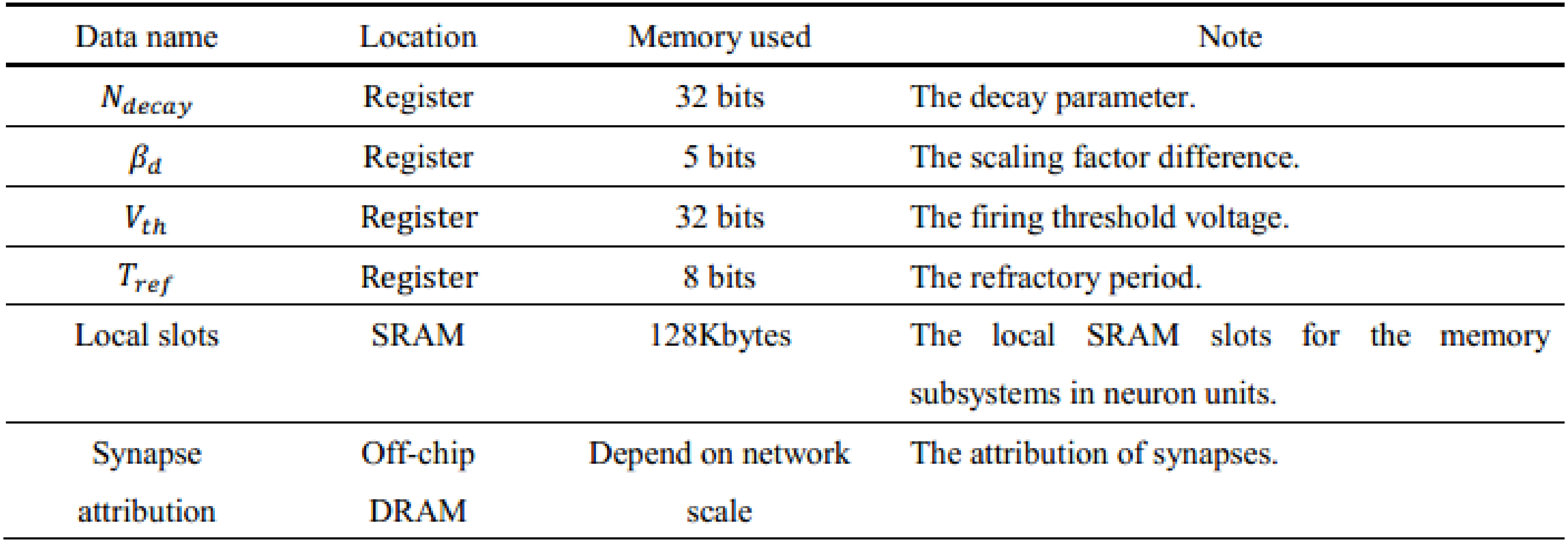
Features:

- 2048 neurons, 15 different synaptic delays, and 2048^2 = 4,194,304 synapses;
- AER format for both input and internal spikes;
- Multiple logical neurons implemented on 8 physical neuron units with time multipexing.



Key Parameters of the Micro-Architecture

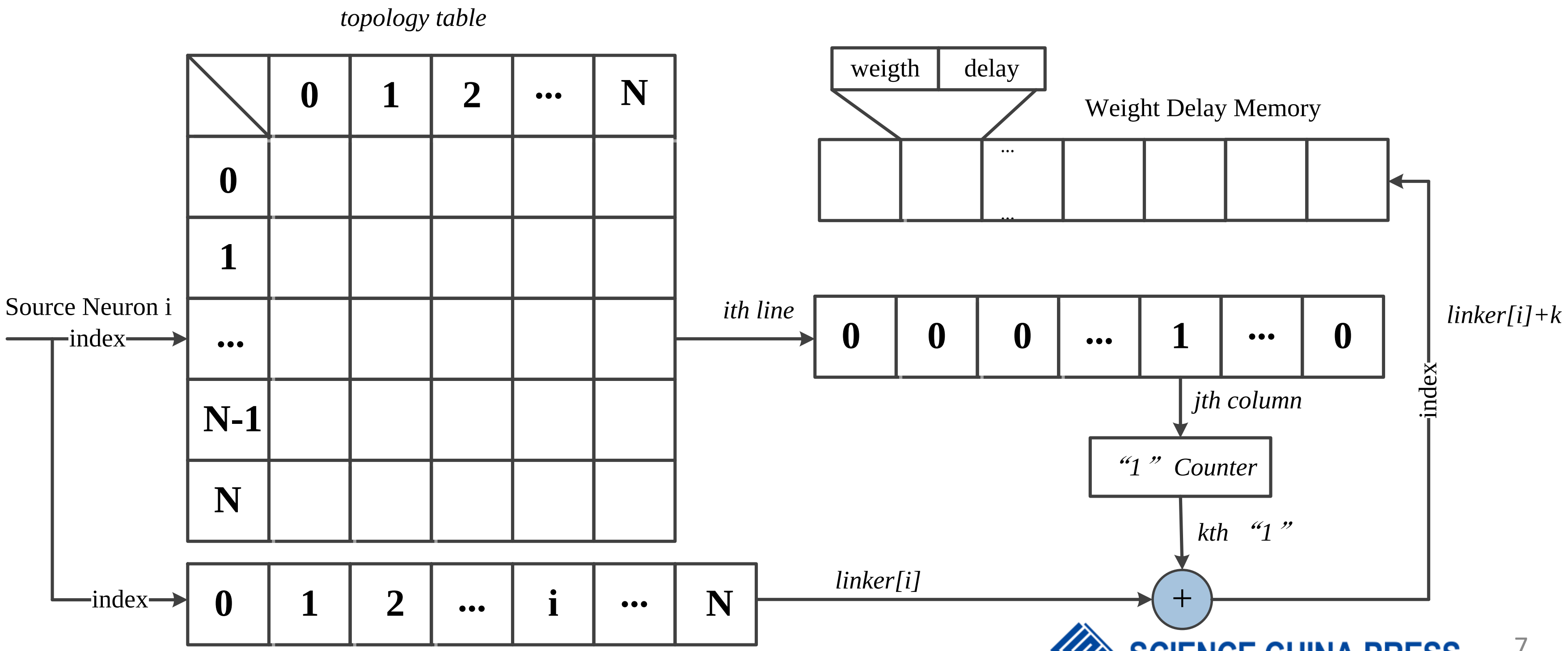
Memory allocation: Use registers, local SRAM, off-chip SDRAM for different types of data.



Data name	Location	Memory used	Note
N_{decay}	Register	32 bits	The decay parameter.
β_d	Register	5 bits	The scaling factor difference.
V_{th}	Register	32 bits	The firing threshold voltage.
T_{ref}	Register	8 bits	The refractory period.
Local slots	SRAM	128Kbytes	The local SRAM slots for the memory subsystems in neuron units.
Synapse attribution	Off-chip DRAM	Depend on network scale	The attribution of synapses.

Data Structure for Storing Synapse Weights in External DRAM

- When the router receives an internal AER with pre-synaptic neuron ID of *i*, the router fetches the index *linker*[*i*] denoting the location of the starting address of the synapse attributes. Then the router reads the *i*th line of the internal topology table. Each synapse attribute starting from the address *linker*[*i*] corresponds to a “1” in the topology table line. If the router detects the *k*th “1” in the *j*th column of line *i*, it sends the synapse word with index (*linker*[*i*] + *k*) to the weight-delay queue of post-synaptic neuron with ID *j*.


topology table

	0	1	2	…	N
0					
1					
…					
N-1					
N					

Source Neuron i

index

index

0	1	2	…	i	…	N

ith line

0	0	0	…	1	…	0

jth column

“1 ” Counter

kth “1 ”

linker[i]

+

index

linker[i]+k

weigth | delay

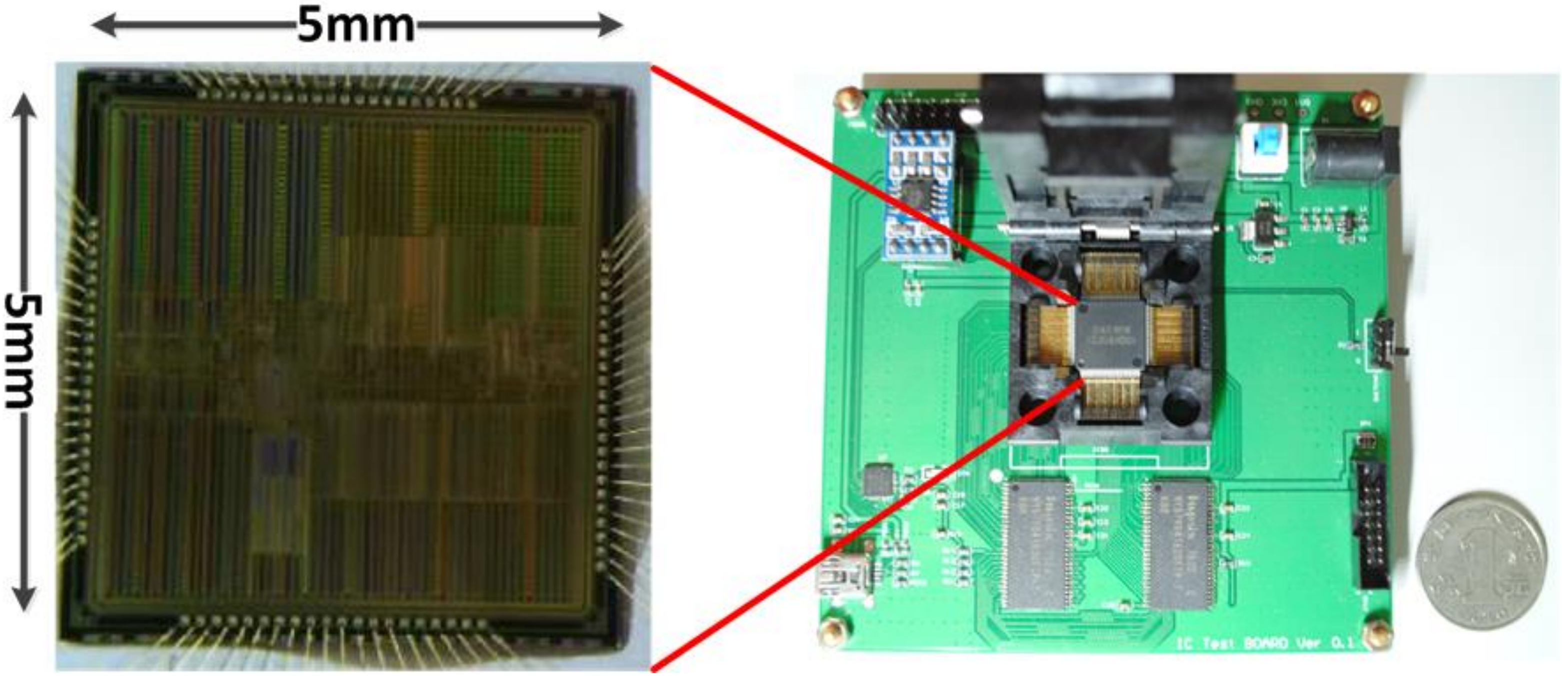
Weight Delay Memory

…

…

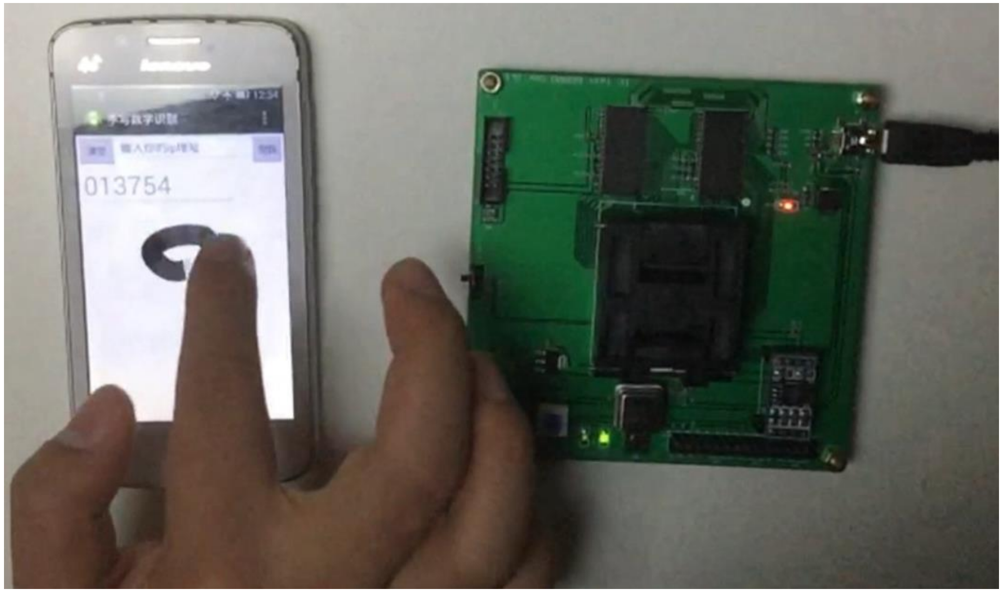
Chip Fabrication

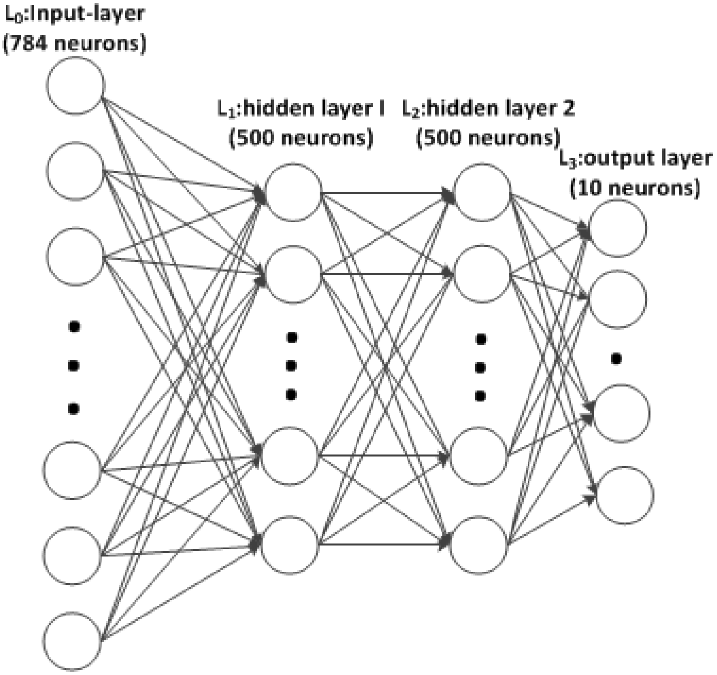
ASIC version of Darwin NPU has been fabricated in SMIC’s 180nm CMOS process, with area of 5x5 mm^2 and 70MHz @worst case. Its power consumption is 0.84mw/MHz for a typical application with 1.8V power supply.



Chip photo and development board

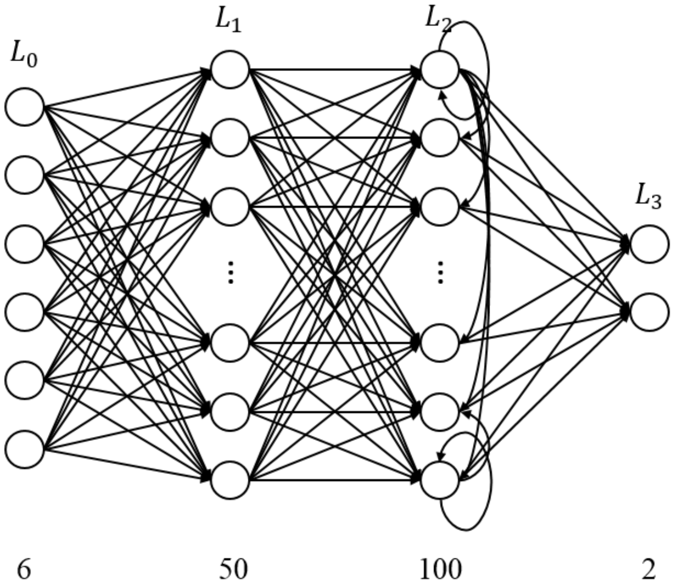
Application Case 1: Handwritten Digit Recognition





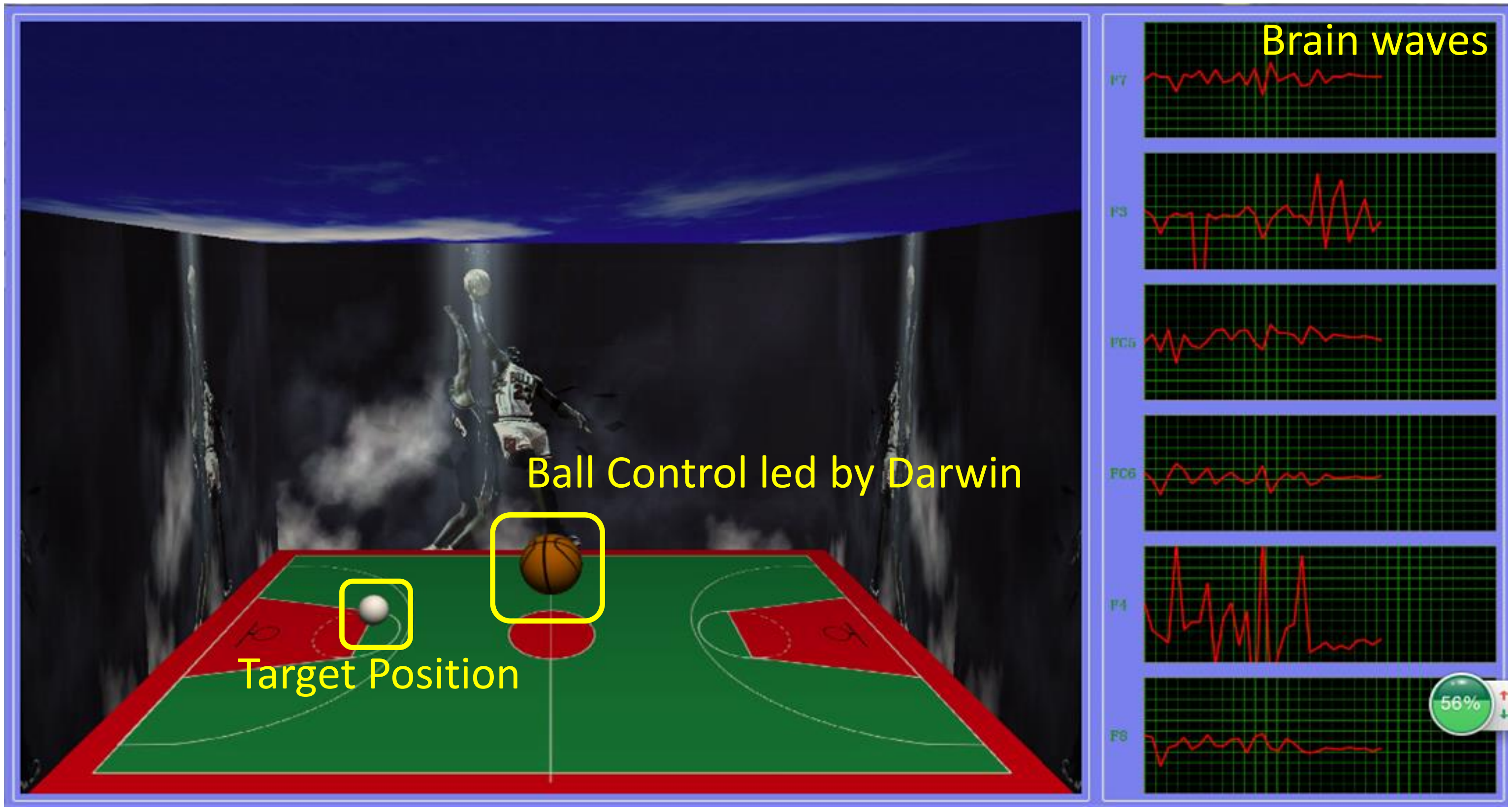
- 4-layer SNN, with full feedforward connection between layers;
- L_0: input layer of 784 neurons;
- L_1 and L_2 : two hidden layers with 500 neurons each;
- L_3: output layer of 10 neurons, each representing a number between 0-9;
- Darwin works at 25MHz for this application .

Application Case 2: EEG Decoding of Motor Imagery



- 4-layer SNN, with full feedforward connection between layers, and recurrent connection within L_2
- L_0: input layer of 6 neurons
- L_1: hidden layer of 50 neurons
- L_2: hidden layer of 100 neurons, with full recurrent connections within the layer
- L_3: output layer of 2 neurons, each representing a binary decision of either left or right imagery motion.
- Darwin works at 25MHz for this application

Application Case 2: EEG Decoding of Motor Imagery



- *Emotiv* headset: sense a user’s brain waves
- *Darwin* NPU: classify whether the user is thinking of left or right, and then control the basketball’s movement to follow the white ball.
- Results: achieve accuracy of **92.7%**. The SNNs is trained by 4000 imagery motion samples, and tested under 4000 real-time captured samples.

Conclusions

- A highly configurable NPU is designed for spiking neural networks;
- Supporting a maximum of 2048 neurons, 15 different synaptic delays and 2048^2 = 4,194,304 synapses;
- The Darwin NPU was fabricated by standard 180nm CMOS technology with area size of 5x5 mm^2 and 70MHz clock frequency @worst case. It consumes 0.84mW/MHz with 1.8 V power supply for typical applications;
- The configurability and efficiency of the hardware is proven by two learning-based classification applications.