SCIENCE CHINA
Information Sciences

• MOO PAPER •

February 2016, Vol. 59 023401:1–023401:5
doi: 10.1007/s11432-015-5511-7

Darwin: a neuromorphic hardware co-processor based on Spiking Neural Networks

Juncheng SHEN[1,3], De MA[2], Zonghua GU[1*], Ming ZHANG[1], Xiaolei ZHU[3], Xiaoqiang XU[1], Qi XU[1], Yangjing SHEN[2] & Gang PAN[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, Hangzhou* 310018*, China;*
[3]*Institute of VLSI Design, Zhejiang University, Hangzhou* 310027*, China*

Received December 4, 2015; accepted December 14, 2015; published online December 23, 2015

Abstract Broadly speaking, the goal of neuromorphic engineering is to build computer systems that mimic the brain. Spiking Neural Network (SNN) is a type of biologically-inspired neural networks that perform information processing based on discrete-time spikes, different from traditional Artificial Neural Network (ANN). Hardware implementation of SNNs is necessary for achieving high-performance and low-power. We present the Darwin Neural Processing Unit (NPU), a neuromorphic hardware co-processor based on SNN implemented with digitallogic, supporting a maximum of 2048 neurons, $2048^2 = 4194304$ synapses, and 15 possible synaptic delays. The Darwin NPU was fabricated by standard 180 nm CMOS technology with an area size of 5×5 mm^2 and 70 MHz clock frequency at the worst case. It consumes 0.84 mW/MHz with 1.8 V power supply for typical applications. Two prototype applications are used to demonstrate the performance and efficiency of the hardware implementation.

Keywords neuromorphic computing, Spiking Neural Networks (SNN), digital VLSI

Citation Shen J C, Ma D, Gu Z H, et al. Darwin: a neuromorphic hardware co-processor based on Spiking Neural Networks. Sci China Inf Sci, 2016, 59(2): 023401, doi: 10.1007/s11432-015-5511-7

1 Introduction

Spiking Neural Network (SNN) is a type of biologically-inspired neural networks that perform information processing based on discrete-time spikes instead of floating point or integer numbers as in traditional Artificial Neural Network (ANN). There are a number of different approaches to developing accelerated implementation of SNN models based on different types of hardware platforms, including (1) Multicore and manycore CPUs, e.g., SpiNNaker Project from the University of Manchester [1]; (2) Graphics Processing Unit (GPU), e.g., CarlSim3 from University of California, Irvine [2]; (3) Digital logic implementation with FPGA or ASIC, e.g., the IBM TrueNorth chip developed in the DARPA SyNAPSE project [3]; (4) Analog and Mixed-Signal implementation, e.g., the ROLLS processor from ETHZ, Switzerland [4].

In this paper, we present the Darwin Neural Processing Unit (NPU), a neuromorphic hardware co-processor based on SNN implemented with digital logic, for use in resource-constrained embedded applications. It has been prototyped on FPGA, and fabricated in SMIC's 180 nm process.

* Corresponding author (email: zonghua@gmail.com)

2 The neuron model

The Leaky Integrate and Fire (LIF) model [5] is a simplified model of biological neuron, widely used in neuromorphic engineering projects. The membrane potential V of a LIF neuron is described by the following equation:

$$C_m \frac{\mathrm{d}V}{\mathrm{d}t} = g_l(V_{\mathrm{rest}} - V) + I, \tag{1}$$

where 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 the membrane potential V rises up to reach the firing threshold V_{th}, a spike (also called an Action Potential) is triggered, and V rapidly rises to a large value, then is reset to $V = V_{\mathrm{reset}}$. Afterwards, there is a refractory period with length of T_{ref}, when the neuron is not responsible to input spikes. At the end of the refractory period, the membrane potential V returns to the resting membrane potential V_{rest}, and starts to be responsive to input spikes again.

To implement the LIF model with digital logic, it is necessary to have a discrete-time version of the LIF model. 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 equations:

$$V_j(t) \leftarrow V_j(t-1)(1 - \Delta t/\tau_m) + \sum_i S_{ij} V_{\max} w_{ij}, \tag{2}$$

$$V_j(t) \leftarrow \begin{cases} 0, & \text{if } t \in [T_f, T_f + T_{\mathrm{ref}}], \\ H(V_{\mathrm{th}} - V_j(t))V_j(t), & \text{otherwise}, \end{cases} \tag{3}$$

$$S_i(t) \leftarrow H(V_i(t) - V_{\mathrm{th}}), \tag{4}$$

where $V_j(t)$ is the membrane potential of neuron j at time step t; the term $\sum_i S_{ij} V_{\max} w_{ij}$ corresponds to the input current I, equal to sum of each input spike current multiplied by the respective synapse weights. Δt is simulation time step size, with a typical value of 0.1 ms; $\tau_m = C_m/g_l$ 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 post-synaptic neuron j; it is positive if the synapse is excitatory; negative if it is inhibitory; V_{th} is the firing threshold; $H(x) = \begin{cases} 1, & x \geqslant 0, \\ 0, & x < 0 \end{cases}$ is the unit step function; V_{rest} and V_{reset} are both assumed to be 0. If the neuron fired an output spike at $t = T_f$, then it remains quiescent for the length of the refractory period during the time interval $[T_f, T_f + T_{\mathrm{ref}}]$, when its membrane potential stays at $V_{\mathrm{reset}} = 0$ and not responsive to input spikes. (The synapse delay does not appear explicitly in (2)–(4), but is modeled as a circular buffer in Figure 1.) The floating-point variables in (2)–(4) are then converted to fixed-point integer variables for implementation with digital logic, with different bit-widths for different variables, e.g., the membrane potential $V_j(t)$ has bit-width of 32, and the synapse weight w_{ij} has bit-width of 16.

3 Implementation details

Figure 2 shows the overall micro architecture of the Darwin NPU. It is designed for power and cost sensitive applications, providing high configurability for both the neuron model and the network topology while minimizing memory cost.

The Address-Event Representation (AER) format is used for information encoding of both input and output spikes. Each spike is represented with an AER packet, which consists of two fields: ID of the neuron that generated the spike and the time stamp when the spike is generated. The NPU works in an event-triggered manner, where most of the NPU logics function only when an input AER packet is received, providing low stand-by power consumption. While the NPU is currently a single-chip system, the standard communication interface defined by the AER format enables future extensions to multi-chip systems interconnected by an AER bus.

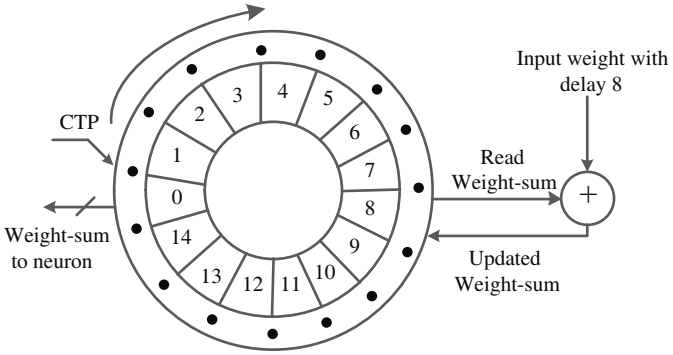

Figure 1 The overall microarchitecture of the NPU.

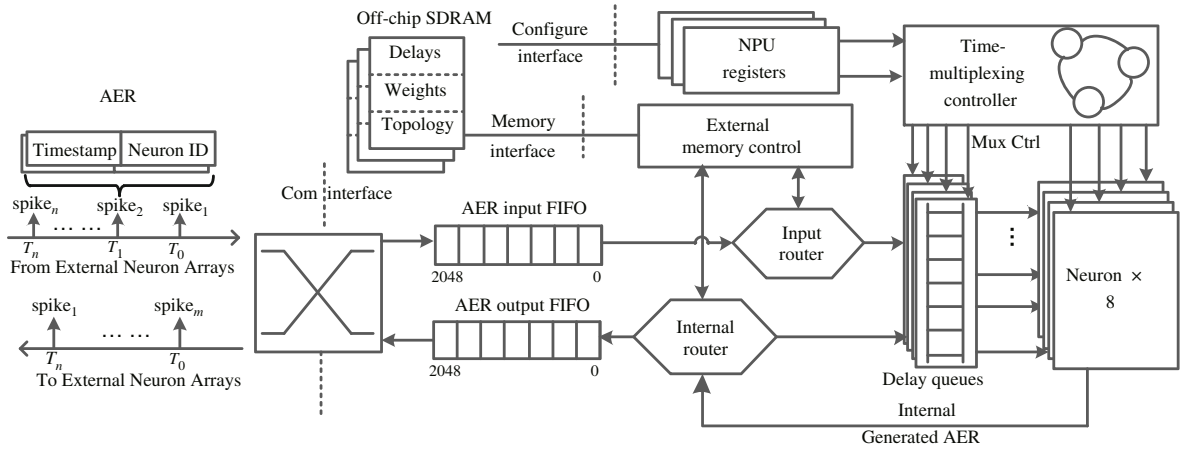

Figure 2 Circular buffer design for the delay queue.

Due to its limited area size, the NPU supports 8 physical neurons on the chip; each neuron can be used to simulate a maximum of 256 logical neurons with time multiplexing, so the whole chip supports a maximum of $8 \times 256 = 2048$ logical neurons.

The main configurable parameters of Darwin NPU include

(1) Global parameters: including all parameters in (2)–(4), which are shared by all neurons (after floating-to-fixed point conversion and parameter consolidation). Since these parameters are used at every time step, they are stored in the on-chip registers to maximize performance, considering their high access frequency and small size.

(2) Per-neuron variables: including each neuron's membrane potential $V_j(t)$, remaining length of the refractory period, and delay queue, stored in an on-chip SRAM of size 16 KB for each physical neuron. Hence the NPU has a total of 8×16 KB $= 128$ KB of on-chip SRAM.

(3) Synapse attributes: including a table of Boolean values encoding the neural network's connection topology, as well as weight and delay attributes of each synapse. Each synapse's attributes are accessed when the synapse is activated by a spike. Its average access frequency is low, since a specific synapse in a large SNN has low average probability of being activated. The synapse attributes can be very large, ranging from several MBs up to GBs depending on the SNN size. Therefore, they are stored in off-chip SDRAM to archive high storage density.

The run time execution consists of the following steps:

1. Spike routing: Each input spike in the form of an AER packet contains a time stamp and source (pre-synaptic) neuron ID, which is used to look up the IDs of destination (post-synaptic) neuron IDs, and the synapse attributes including their weights and delays stored in off-chip DRAM.

2. Synapse delay management: Each synapse has a configurable delay parameter, i.e., the delay from spike generation of the pre-synaptic neuron to spike reception of the post-synaptic neuron. As shown in

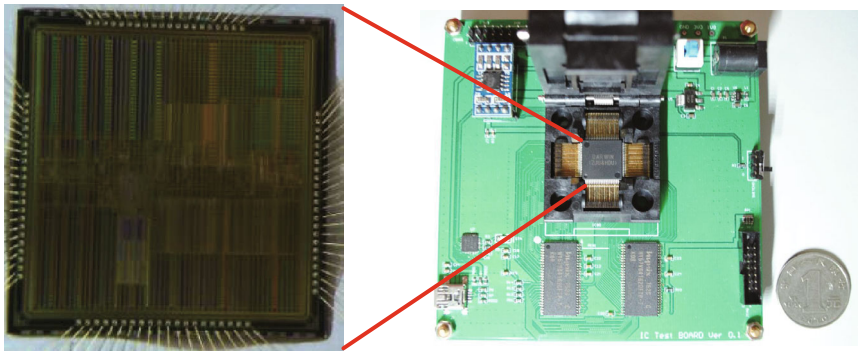

Figure 3 Photos of the die and the demonstration PCB board.

Figure 1, each neuron is associated with a delay queue consisting of 15 slots implemented as a circular buffer; the kth slot stores the sum of input synapse weights (used to compute the term $\sum_i S_{ij} V_{\max} w_{ij}$ in (2)) with synaptic delay of k time steps. At each time step, the Current Time-step Pointer (CTP) is shifted by one slot, and the content of the slot previously pointed to by CTP is sent to the neuron as synaptic input.

3. Neuron status update: Each neuron performs status update based on (2)–(4). If an output spike is generated, it is sent to the spike router (input or internal router in Figure 2) in the form of AER packets.

4 Experiment results

The Darwin NPU has been implemented with synthesizable Verilog, and fabricated with standard 180 nm CMOS technology with an area size of 5×5 mm^2 and 70 MHz clock frequency at the worst case. It consumes 0.84 mW/MHz with 1.8 V power supply for typical applications. It has been integrated with a RISC CPU to form a complete System-on-Chip (SoC), based on the open-source OpenCores minsoc project (http://opencores.org/project, minsoc). In addition to the NPU and CPU, the SoC also consists of a local bus, 64 KB SRAM, SPI flash, UART controller and SDRAM controller. We perform protocol conversion between UART and USB, so that the SoC can be used as an USB device and attached to a host PC. Figure 3 shows the die photo, and the prototype PCB board.

Next, we present two application case studies.

(1) Application case 1: handwritten digit recognition.

The first application case study is the Spiking Deep Belief Network (DBN) from [6] for handwritten digit recognition. It is a 4-layer SNN, with full feed forward connection between layers. The input layer consists of 784 neurons, each representing an image pixel in a 28 × 28 pixel image; each input neuron emits a spike train that uses firing rate coding to encode pixel intensity. Each of the two hidden layers consists of 500 neurons; and the output layer consists of 10 neurons, each representing a digit of 0–9. The output neuron with the highest firing rate is selected as the classification output.

We consider another SNN hardware accelerator Minitaur [6], which implemented the same Spiking DBN on a Xilinx Spartan-6 FPGA. The Darwin NPU is configured to have clock speed of 25 MHz, with average latency of 0.16 s for recognizing each digit, and overall classification accuracy of 93.8%. In comparison, Minitaur has clock speed of 75 MHz, average latency of 0.152 s, and classification accuracy of 92%. The Darwin NPU achieved similar average latency with a lower clock speed, and slightly better classification accuracy.

(2) Application case 2: EEG decoding of motor imagery.

The second application case study is decoding of Electroencephalogram (EEG) signals. We use the Emotiv headset to collect EEG signals for real-time classification of the user's motor imagery, i.e., whether the user is thinking of left or right direction, and use the result to control the movement of a virtual ball on the screen. The SNN has 4 layers, with full feed forward connection between layers. The input layer consists of 6 neurons, each representing an EEG channel; each input neuron emits a spike train that uses

firing rate coding to encode EEG signal amplitude. The first hidden layer consists of 50 neurons, and the second hidden layer consists of 100 neurons, with full recurrent connections within the layer; the output layer consists of 2 neurons, each representing a binary decision of either left or right motor imagery. The output neuron with the highest firing rate is selected as the classification output. The training set consists of 4000 EEG signal fragments, and the test set consists of 4000 EEG signal fragments captured and decoded in real-time. The classification accuracy is 92.7%.

5 Conclusion and future work

We present the Darwin NPU, a neuromorphic hardware co-processor based on SNN, supporting a maximum of 2048 neurons, $2048^2 = 4194304$ synapses, and 15 possible synaptic delays. As part of future work, we plan to use it as a Processing Element in a Network-on-Chip (NoC) architecture, using the AER format for input and output spikes, in order to scale up the SNN size to potentially millions of neurons instead of mere thousands.

Acknowledgements
This work was supported by National Basic Research Program of China (973 Program) (Grant No. 2013CB3295-04).

Conflict of interest The authors declare that they have no conflict of interest.

Supporting information
The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Furber S B, Galluppi F, Temple S, et al. The spinnaker project. Proc IEEE, 2014, 102: 652–665
2 Beyeler M, Carlson K D, Chou T S, et al. CARLsim 3: a user-friendly and highly optimized library for the creation of neurobiologically detailed spiking neural networks. In: Proceedings of International Joint Conference on Neural Networks (IJCNN), Killarney, 2015. 1–8
3 Merolla P A, Arthur J V, Alvarez-Icaza R, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. Science, 2014, 345: 668–673
4 Qiao N, Mostafa H, Corradi F, et al. A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128 K synapses. Front Neurosci, 2015, 9: 141
5 Dayan P, Abbott L F. Theoretical Neuroscience. Cambridge: MIT Press, 2001. 11–52
6 Neil D, Liu S C. Minitaur, an event-driven FPGA-based spiking network accelerator. IEEE Trans Very Large Scale Integr Syst, 2014, 22: 2621–2628